Received: 17 July 2022 | Revised: 9 October 2022 | Accepted: 14 October 2022

DOI: 10.1111/cas.15629

ORIGINAL ARTICLE

Cancer Science WILEY

Serum untargeted metabolomics reveal metabolic alteration of non-small cell lung cancer and refine disease detection

Jiaoyuan Li | Ke Liu | Zhi Ji | Yi Wang | Tongxin Yin | Tingting Long | Ying Shen | Liming Cheng

Department of Laboratory Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Correspondence
Ying Shen and Liming Cheng, Department of Laboratory Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China.
Emails: sying830@163.com; chengliming2015@163.com

Funding information
National Key Research and Development Plan Program of China, Grant/Award Number: 2016YFC1302702; National Natural Science Foundation of China, Grant/Award Number: 81572071 and 81903394

Abstract

This study was performed to characterize the metabolic alteration of non–small-cell lung cancer (NSCLC) and discover blood-based metabolic biomarkers relevant to lung cancer detection. An untargeted metabolomics-based approach was applied in a case–control study with 193 NSCLC patients and 243 healthy controls. Serum metabolomics were determined by using an ultra high performance liquid chromatography–tandem mass spectrometry (UHPLC-MS/MS) method. We screened differential metabolites based on univariate and multivariate analysis, followed by identification of the metabolites and related pathways. For NSCLC detection, machine learning was employed to develop and validate the model based on the altered serum metabolite features. The serum metabolic pattern of NSCLC was definitely different from the healthy condition. In total, 278 altered features were found in the serum of NSCLC patients comparing with healthy people. About one-fifth of the abundant differential features were identified successfully. The altered metabolites were enriched in metabolic pathways such as phenylalanine metabolism, linoleic acid metabolism, and biosynthesis of bile acids. We demonstrated a panel of 10 metabolic biomarkers which representing excellent discriminating capability for NSCLC discrimination, with a combined area under the curve (AUC) in the validation set of 0.95 (95% CI: 0.91–0.98). Moreover, this model showed a desirable performance for the detection of NSCLC at an early stage (AUC = 0.95, 95% CI: 0.92–0.97). Our study offers a perspective on NSCLC metabolic alteration. The finding of the biomarkers might shed light on the clinical detection of lung cancer, especially for those cancers in an early stage in Chinese population.

KEYWORDS
biomarker, diagnosis, machine learning, metabolomics, non–small-cell lung cancer

1 | INTRODUCTION

Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer death worldwide, with an ~11.4% of cancer cases and a mortality up to 18.0% of the total cancer deaths in 2020.[1] The survival of patients with lung cancer at 5 years after diagnosis is only 10%–20% in most countries.[2] Early screening and detection using biopsy and low dose computed tomography (LDCT) methods can greatly reduce the lung cancer mortality by 24% in men and 33% in women, and substantially bring down the personal and financial burden.[3] However, conventional diagnosis of lung cancer is still limited considering the insufficient accuracy, invasiveness, high cost and low throughput. Improved means for identifying individuals with high risk of lung cancer are thus urgently needed in order to enhance patient care and optimize disease management.

In recent years, liquid assays have generated extensive attention in disease diagnosis areas as they are noninvasive, easy in sampling and low in cost. Utilizing blood-based assays, many genomic and proteinic biomarkers have been newly identified, although their sensitivities or specificities were far from satisfied.[4–6] Compared with those genomic, transcriptomics, or proteinic biomarkers, metabolic indicators are directly involved in tumor metabolism under the influence of both genetic modification and environmental stress, thus providing more distal information on cancer initiation and progression.[7] Changes of some metabolites have been reported to be associated with complex diseases including lung cancer, suggesting a potential role of metabolites in precise diagnosis of common diseases. Recent advances in technology allow for the detection of more small molecule metabolites, enabling abnormal metabolites screening in relatively high throughput. In leveraging these technological developments, researchers have identified a small subset of serum or plasma metabolites as diagnosis biomarkers of multiple cancers such as colorectal cancer, hepatocellular carcinoma, breast cancer, and pancreatic cancer.[8–11] For example, a combination of metabolic biomarkers (creatine, inosine, beta-sitosterol, sphinganine, and glycocholic acid) have demonstrated higher accuracy and specificity in the diagnosis of pancreatic cancer than conventional biomarkers (CA125, CA19-9, CA242, and CEA).[11] For lung cancer, several metabolic biomarker panels or patterns have also been developed, enabling the discrimination between lung cancer patients and healthy individuals.[12–14] However, these subsets of metabolites were usually screened from a class of known metabolites such as targeted metabolomics, lipidomics, or amino acid profiles, the untargeted metabolome-based approach has rarely been reported.

To fully recognize the altered metabolites and metabolic pathways that regulate tumor initiation and progression, as well as to characterize sensitive and specific biomarkers for precise diagnosis of non–small-cell lung cancer (NSCLC) in Chinese population, we conducted an untargeted metabolomics-based case–control study in this work. Serum metabolomics of 193 NSCLC patients and 243 healthy controls were performed using an ultra performance liquid chromatography–tandem mass spectrometry (UPLC-MS/MS) method. The differential metabolites between NSCLC patients and healthy controls were screened and identified, followed by the characterization of related metabolic pathways. For NSCLC detection, the subjects were randomly assigned into training set and validation set, and machine learning was employed to develop and validate the model based on the altered serum metabolite features.

2 | METHODS

2.1 | Study subjects

We enrolled 193 NSCLC patients and 243 healthy individuals for serum untargeted metabolome analysis, 70% of which (135 NSCLC cases and 170 controls) were randomly assigned into a training set and the remaining 30% subjects (58 NSCLC cases and 73 controls) were assigned into a validation set. All the subjects were consecutively recruited from October 2016 to December 2019 at Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology (HUST), Wuhan, China. NSCLC patients were newly diagnosed with histological or cytological confirmation, among which patients with previous surgery, chemotherapy or radiotherapy were excluded. In addition, metastatic NSCLC, familial lung cancer, recurrent cancer, or multiple primary tumors were also excluded. The TNM staging of NSCLC was defined according to the eighth edition of the TNM cancer staging system from The American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control (IUCC). Cancer-free controls were randomly selected from healthy individuals who visited the health check-up center in the same hospital during the same period of NSCLC patient recruitment. All the healthy controls were frequency matched to NSCLC cases with sex and age (±5 years). The basic demographic characteristics of the participants including sex, age, smoking, and drinking status were collected through medical records or face-to-face interview. This study was approved by the Institutional Review Committee of the Tongji Medical College, HUST.

2.2 | Sample preparation

Peripheral blood sample was acquired from each participant through intravenous collection at early morning after overnight fasting, and centrifuged immediately at 3000 rpm for 5 min. The upper layer of serum was then collected and stored in a −80°C freezer immediately until use. Frozen serum samples were thawed at 4°C and 140 μl of the supernatants were transferred into a 1.5 ml tube, mixed with 420 μl of ice-cold methanol to precipitate the proteins. The mixture was vortexed for 20 s and centrifuged at 14,000 g for 10 min, and then 350 μl of the supernatants were transferred into a new tube, dried under nitrogen at 40°C and reconstituted with 70 μl water:methanol mixture (50%:50%, v:v). After vortexing (20 s) and centrifugation (14,000 *g* for 10 min), the supernatants were then used to perform untargeted metabolomics analysis or targeted assessment. In addition, a serum pool was prepared by mixing equal volume of the individual samples, which was used for quality control (QC).

2.3 | Untargeted metabolomics

LC-MS/MS analysis was performed on a UPLC Ultimate 3000 system coupled to a Q Exactive™ mass spectrometer (Thermo Fisher Scientific, Bremen, Germany), and operated in the positive and negative electrospray ionization modes (one run for each mode). The heated electrospray ionization (HESI) source was used for both modes, and a spray voltage 3.8 kV for positive mode and 3.2 kV for negative mode, capillary temperature of 320°C, sheath gas flow of 40 arbitrary units (AU), auxiliary gas flow of 10 AU, spare gas flow of 0 AU, and S-lens random forest (RF) level of 60 V. During the full-scan acquisition, which ranged from 70 to 1000 m/z, the instrument operated at 70,000 resolution (m/z = 200), with an automatic gain control (AGC) target of 3e6 charges and a maximum injection time (IT) of 200 ms. For MS^2 analyses, the isolation window was set at 0.4 m/z, and the instrument was operated at 17,500 resolution (m/z = 200) with an AGC target of 1e5 charges and maximum IT of 50 ms. The stepped normalized collision energy (NCE) was set at 20, 40, or 60 eV. The system was controlled using Xcalibur 2.2 software (Thermo Fisher Scientific).

For chromatographic separation, a Thermo Scientific Hypersal GOLD™ C18 column (2.1 mm × 100 mm, 1.9 μm) was used. The mobile phase composed of mobile phase A of 0.1% formic acid in water and mobile phase B consisting of 0.1% formic acid in acetonitrile. A multistep gradient was as follows: 0–1 min, 98% A; 1–9 min, from 98%A to 98%B; 9–12 min, 98%B; 12–15 min, 98%A. The gradient was operated at a flow rate of 0.4 ml/min over a run time of 15 min for both the negative and the positive modes. The column temperature was 40°C. The ultra high performance liquid chromatography (UHPLC) autosampler temperature was set at 4°C and the injection volume for each sample was 5 μl. The case and control samples were analyzed alternately. Three blanks and four QC samples were injected to equilibrate the system before each analytic series, and one QC sample was injected after every 10 samples to monitor the reproducibility of the LC-MS/MS analysis.

2.4 | Data processing and differential metabolites screening

The raw data for metabolomics were processed using Thermo Scientific Compound Discover™ 2.1 software (Thermo Scientific). If the quantitative value of a peak signal was less than five times that of the median of blank samples in more than 75% of samples, the detection rate of this signal was considered low and then removed. Peak signals with variation >30% in QC samples were also removed. The metabolomes data for all the samples were downloaded into an MS Excel® file for further analysis.

We performed both multivariate and univariate analysis to identify the differential metabolites in serum of NSCLC patients compared with healthy controls. The multivariate analyses were achieved through supervised partial least-squares discriminant analysis (PLS-DA) and orthogonal partial least square-discriminant analysis (OPLS-DA) via SIMCA® (Umetrics), while the univariate analyses were accomplished by Mann–Whitney *U*-test and volcano plot using open-access software R 3.5.3 (https://www.r-project.org/). MS data were normalized with the total sum of all detected ions, centered and scaled using Pareto scaling before analysis. A metabolite was considered as significantly altered metabolite if the value of variable importance in the projection (VIP) > 1 and $p < 0.05$ in multivariate analyses, and the fold change >1.2 or <0.8 with false discovery rate (FDR) *q*-value <0.05 in univariate analyses. Next, we sorted the changed metabolites by their relative abundances in serum and the top 100 metabolites were further identified by matching the acquired MS/MS data against data in mzCloud database, and a prediction score of ≥ 70 was considered acceptable. To validate the accuracy of prediction, we further randomly selected 10 of the identified metabolites and confirmed them by the commercial chemical standards (Sigma-Aldrich, Merck KGaA). The identified metabolites were mapped into biochemical pathways through metabolic pathway and enrichment analyses based on the KEGG database on MetaboAnalyst (https://www.metaboanalyst.ca/).

2.5 | Machine learning

Eight machine learning approaches, including decision tree, random forest (RF), support vector machine (SVM), logistic regression, XGBoost, linear support vector classification (SVC), stochastic gradient descent (SGD) and linear discriminant analysis (LDA) were employed for NSCLC detection. The diagnostic model was established based on the features of biomarker candidates in the training set, and internally validated in the validation set. All the identified metabolites were selected as promising diagnostic biomarkers for further machine leaning study. Taking both discriminating capacity and clinical convenience into consideration, we constructed the machine learning models with the top important 5–12 metabolites by adding one variable at a time, and chose the panel with the best discriminating ability and least number of variables as our optimal model. A receiver operating characteristic (ROC) curve was utilized to evaluate the diagnostic performance of the models.

All the machine learning analyses were performed on the Deepwise & Beckman Coulter DxAI platform (http://dxonline.deepwise.com), and the visualization of the results was achieved by using GraphPad Prism 8.0 (GraphPad Software).

3 | RESULTS

3.1 | Basic characteristics of study subjects

We recruited 193 NSCLC cases and 243 cancer-free controls in our untargeted metabolomics study. There were no significant differences about the sex and age between cases and controls ($p = 0.358$ for sex and $p = 0.378$ for age). About 52.6% of the cases were smokers (current or ever), which was significantly higher than that in the control group (39.9%, $p = 0.016$). The percentages of drinker were 34.4% in the cases and 37.8% in the controls ($p = 0.410$). Histological

or cytological examination confirmed that 56.0% of the tumors were lung adenocarcinomas (LUAD) and 29.0% were lung squamous cell carcinomas (LUSC). All the remaining tumors were other non–small-cell histological types or NSCLC which could not be clearly classified. Based on the TNM staging criteria, 38.8% and 38.3% of the total NSCLC patients in metabolomics study were diagnosed in an early stage (0, I, or II stages) and late stage (III and IV stages), respectively. The demographic and clinical characteristics of the study subjects are presented in Table 1.

3.2 | Comparison of metabolic profiles in NSCLC and healthy controls

From UPLC-MS/MS analyses, 940 metabolite features in positive ion mode and 1904 metabolite features in negative ion mode were captured and selected for subsequent analyses. The PLS-DA analysis optimized six principal components (PCs) and eight PCs for model fitting in positive ion mode and negative ion mode, respectively. The permutation test ($n = 100$) showed that the constructed models had acceptable reliability and predictability with no overfitting in both the positive and negative ion modes (Figure S1A,B). As shown in the PC plots of OPLS-DA analysis (Figure 1A,B), the NSCLC patients and healthy controls were basically separated based on the metabolite features, indicating an essential difference of the metabolism between NSCLC patients and healthy individuals. According to our criteria, in total 381 metabolite features, including 145 metabolites in positive ion mode and 236 features in negative ion mode, were found significantly altered in NSCLC compared with healthy control in multivariate analyses. In addition, univariate analyses identified 380 differential metabolite features in the positive ion mode (Figure 1C) and 745 differential ones in the negative ion mode (Figure 1D). The results of multivariate and univariate analyses were intersected, and in total 278 metabolite features (Figure 1E), including 114 ones in positive ion mode and 164 ones in negative ion mode, were finally regarded as the differential metabolites which were significantly altered in the serum of NSCLC patients compared with healthy people. Among these peaks, 128 features (77 ones in positive mode and 51 ones in negative mode) were upregulated and 150 features (37 ones in positive mode and 113 ones in negative mode) were downregulated in the serum of NSCLC patients (Figure 1E).

TABLE 1 The basic information about study subjects

	Cases (%)	Controls (%)	*p* value
Sex			
Male	132 (68.4%)	176 (72.4%)	0.358
Female	61 (31.6%)	67 (27.6%)	
Age	59.3 (7.6)	60.0 (9.3)	0.378
Smoking[a]			
No	91 (47.4%)	146 (60.1%)	0.016
Yes	101 (52.6%)	97 (39.9%)	
Drinking[b]			
No	126 (65.6%)	150 (62.2%)	0.41
Yes	66 (34.4%)	91 (37.8%)	
Pathologic type			
LUAD	108 (56.0%)	–	
LUSC	56 (29.0%)	–	
Others	29 (15.0%)	–	
Clinical stage			
0+I	40 (20.7%)	–	
II	35(18.1%)	–	
III	35 (18.1%)	–	
IV	39 (20.2%)	–	
Unknown	44 (22.8%)	–	

Abbreviations: LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

[a]Smoking information of one NSCLC case was missing.

[b]Drinking information of one NSCLC case and two controls was missing.

3.3 | Identification of altered metabolites

As mentioned above, in total 278 features were significantly altered when comparing the metabolic profiles of the serum of NSCLC patients and healthy subjects. Considering the diversity of their content in serum, we sorted these metabolites by their relative abundances and only identified the top 100 abundant metabolites. After database retrieval, 21 of them were successfully identified by MS prediction using mzCloud. To verify the accuracy of the prediction, we randomly selected 10 of the metabolites and confirmed them by chemical standards. As a result, all the retention times of the metabolites were repeated accurately on our technique. The detailed information about the 21 identified metabolites is presented in Table 2. Among these, 12 metabolites were upregulated, while the remaining nine metabolites were downregulated in the serum of NSCLC patients. As shown in Figure 1F, the expression pattern of these 21 metabolites in serum rendered obvious distinction between NSCLC cases and healthy controls. Enrichment analysis showed that the most significantly enriched pathway was phenylalanine metabolism, as two of the 21 metabolites, containing L-phenylalanine and hippuric acid, hit the pathway (Figure 1G). Linoleic acid metabolism was also among the top enriched pathways. Interestingly, we found that bile acids, including glycocholic acid (GCA), cholic acid (CA) and glycoursodeoxycholic acid (GUDCA), were upregulated in our NSCLC patients, contributing to the enrichment of bile acid metabolism.

3.4 | Discovery of diagnostic model using candidate metabolites

Through machine learning, the 21 identified metabolites were ordered by importance and the top 5–12 important metabolites were selected as components of the candidate diagnosis models. As

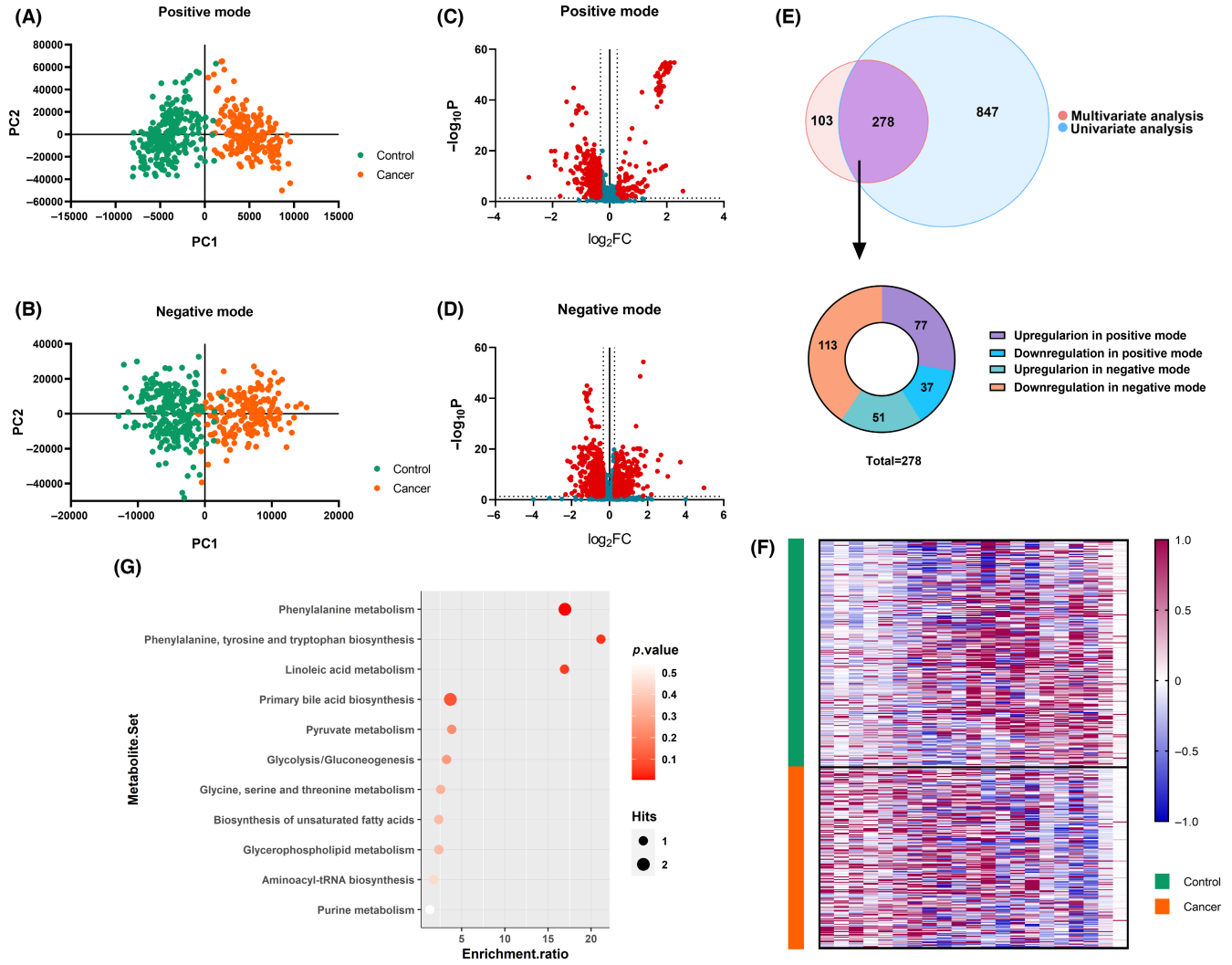

FIGURE 1 Analyses of serum metabolic profiling in non–small-cell lung cancer (NSCLC) patients and healthy controls and screening of metabolites relevant to lung cancer. (A) Orthogonal partial least-squares discriminant analysis (OPLS-DA) of metabolites in positive mode. (B) OPLS-DA of metabolites in negative mode. (C) Volcano plot of metabolites by univariate analysis in positive mode. (D) Volcano plot of metabolites by univariate analysis in negative mode. (E) Venn diagram displaying the screening and constitution of altered metabolites in the serum of NSCLC compared with healthy people. (F) Expression heatmap of the 21 identified metabolites in the serum of NSCLC patients and healthy controls. (G) Enriched metabolic pathways of differential metabolites.

shown in Table S1, the model with the top 10 metabolites presented the highest AUC and the least number of variables. We thus chose this panel of metabolites in the following analyses. The 10 metabolites included in this diagnostic model were hypoxanthine, linoleic acid, 2,4-dihydroxybenzoic acid, 11,12-epoxy-(5Z,8Z,11Z)-icosatrienoic acid, 16-hydroxyhexadecanoic acid, testosterone sulfate, choline, piperine, CA, and GUDCA. The detailed performances of this panel of biomarkers are shown in Table 3, all the models, expect decision tree, showed close and admirable capacity for NSCLC detection in the validation set, with area under the curve (AUC) ranging from 0.93 to 0.95. By applying this panel of biomarkers using LDA method, the AUCs for NSCLC discrimination reached 0.93 (95% CI: 0.91–0.96) and 0.95 (95% CI: 0.91–0.98) in the training set and validation set, respectively (Figure 2A). The predicted score for each subject in the validation set is presented in Figure 2B. We next compared the performance of the single metabolites and the panel containing all the components using all the study subjects in both the training and validation sets. The results demonstrated that the AUCs ranged from 0.63 to 0.77 for single metabolites, while the combination of 10 metabolites greatly improved the AUC to 0.95 (Figure 2C). Under this model, the sensitivity and specificity for NSCLC reached 85.0% and 88.5%, respectively. In addition, we found that the predicted scores of NSCLC patients were not significantly altered in different pathologic types (LUAD or LUSC) and different stages (early stage or advanced stage), all of which were significantly higher than the scores in healthy controls (Figure 2D). In line with this, ROC curves in Figure 2E demonstrated the dramatical diagnostic performances for the discriminant model in both LUAD (AUC = 0.93, 95% CI: 0.90–0.96) and LUSC (AUC = 0.94, 95% CI: 0.91–0.98). It is worth nothing that the panel of metabolic biomarkers also exhibited

TABLE 2 Detailed information about the 21 identified metabolites

Name	Formula	VIP	Fold change	*p*-adjusted	MW	RT (min)	mzCloud score
11,12-Epoxy-(5Z,8Z,11Z)-icosatrienoic acid	$C_{20}H_{32}O_3$	4.0	3.04	1.09E−20	320.23456	9.04	77.9
Cholic acid	$C_{24}H_{40}O_5$	3.5	2.85	2.37E−06	408.28694	7.30	73.7
11-Deoxyprostaglandin	$C_{20}H_{34}O_4$	3.95	2.4	9.46E−21	338.24517	9.04	82.2
Glycocholic acid	$C_{26}H_{43}NO_6$	2.6	2.10	7.83E−10	465.30825	6.72	78.0
Docosahexaenoic acid ethyl ester	$C_{24}H_{36}O_2$	2.1	2.04	9.45E−05	356.2702	8.16	85.6
Glycoursodeoxycholic acid	$C_{26}H_{43}NO_5$	5.8	1.83	1.48E−13	449.31357	7.40	83.0
Hypoxanthine	$C_5H_4N_4O$	5.3	1.33	1.09E−13	136.03814	1.39	90.8
L-(+)-Lactic acid	$C_3H_6O_3$	10.4	1.32	4.46E−18	90.03156	1.21	98.7
α-Aspartylphenylalanine	$C_{13}H_{16}N_2O_5$	1.2	1.25	1.18E−04	280.10546	4.14	93.3
6-Hydroxycaproic acid	$C_6H_{12}O_3$	1.0	1.24	1.16E−05	132.07842	4.82	90.7
Benzoic acid	$C_7H_6O_2$	1.6	1.23	2.96E−05	122.03657	4.16	90.3
L-Phenylalanine	$C_9H_{11}NO_2$	1.2	1.22	4.02E−10	165.07872	3.27	91.3
16-Hydroxyhexadecanoic acid	$C_{16}H_{32}O_3$	2.7	0.79	9.92E−08	272.23472	10.10	90.7
Testosterone sulfate	$C_{19}H_{28}O_5S$	7.9	0.77	1.22E−05	368.16508	6.93	90.3
Choline	$C_5H_{13}NO$	7.5	0.77	1.76E−16	103.09934	1.01	88.7
Hippuric acid	$C_9H_9NO_3$	3.3	0.72	1.36E−04	179.05791	4.55	86.6
Acetyl-β-methylcholine	$C_8H_{17}NO_2$	1.9	0.67	1.70E−13	159.12548	1.08	86.0
Linoleic acid	$C_{18}H_{32}O_2$	2.3	0.64	9.46E−21	298.25025	10.38	75.4
2,4-Dihydroxybenzoic acid	$C_7H_6O_4$	3.2	0.56	1.20E−13	154.02634	4.90	83.0
Salicylic acid	$C_7H_6O_3$	1.8	0.35	6.66E−08	138.03141	5.81	97.2
Piperine	$C_{17}H_{19}NO_3$	3.3	0.31	2.00E−13	285.13539	7.66	93.7

Abbreviations: MW, molecular weight; RT, retention time; VIP, variable importance in the projection.

marked capacity for NSCLC detection with different clinical stages, especially for early-stage NSCLC (AUC = 0.95, 95% CI: 0.92–0.97). These results indicated that the discriminant model is promising for the clinical diagnosis of NSCLC patients against healthy controls.

4 | DISCUSSION

Metabolomics is an emerging discipline that monitors intermediates and products of cellular metabolism, which can be influenced by both xenobiotics and exogenous factors, and contribute to cellular function and dysfunction. Hence, the metabolome provides a phenotypic evaluation of cellular and systemic health, with potential implications for disease pathogenesis, biomarker discovery, drug effectiveness, and personalized medicine.[15] In the present study, in order to characterize the metabolic pattern of NSCLC, we compared the serum metabolites of NSCLC patients and healthy subjects by utilizing an untargeted metabolomics profile. Through integrated screening based on both univariate and multivariate analysis, we found 278 altered features in serum of NSCLC patients, and one-fifth of the abundant metabolites were successfully identified by matching the mzCloud database. The altered metabolites were enriched in pathways involving phenylalanine metabolism, linoleic acid metabolism, and the biosynthesis of bile acids. Moreover, a model comprising 10 metabolic biomarkers was demonstrated to have excellent discriminating capability (AUC = 0.95) for NSCLC classification from healthy people, suggesting the tremendous potential of metabolite panels for lung cancer diagnosis.

The abnormality of metabolites can aid in the illumination of tumor pathological mechanisms since it is the direct manifestation of biological dysfunctions, as the downstream of genome, transcriptome, and proteome. In line with plenty of prior researches, our study confirmed that the circulating metabolic pattern of NSCLC is definitely different with that of healthy condition. We identified 21 metabolites from the top 100 abundant altered metabolic features in serum of NSCLC, some of which have been reported in lung cancer or other tumors. For example, we found lactic acid was elevated in serum of NSCLC patients comparing with healthy subjects. This result is consistent with a previous study reporting serum NSCLC metabolomics based on nuclear magnetic resonance (NMR) fingerprints,[16] and supported by the finding of increased lactate level in lung tumors compared with surrounding normal tissues.[17,18] Lactate is a tricarboxylic acid (TCA) cycle carbon source for NSCLC and serves as a fuel for tumor cells in vivo.[19,20] The elevated lactate level thus indicates an active glycolysis processing and the unique Warburg effect in the energy metabolism of tumors. A significant reduction in linoleic acid was observed in the serum of NSCLC patients in our study, in accordance with a previous report

TABLE 3 The performance of the 10-marker panel for NSCLC detection using eight common machine learning algorithms

Algorithm	Training set						Validation set					
	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Decision tree	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.79	0.79	0.78	0.74	0.83
Random forest	1.00	0.99	0.97	1.00	1.00	0.98	0.93	0.84	0.74	0.92	0.88	0.82
SVM	0.96	0.91	0.88	0.93	0.91	0.91	0.94	0.86	0.83	0.89	0.86	0.87
Logistic regression	0.94	0.88	0.87	0.89	0.86	0.90	0.94	0.89	0.88	0.89	0.86	0.90
XGBoost	1.00	0.99	0.97	1.00	1.00	0.98	0.94	0.86	0.84	0.88	0.84	0.88
LinearSVC	0.94	0.88	0.86	0.89	0.86	0.89	0.94	0.89	0.88	0.89	0.86	0.90
SGD	0.91	0.83	0.70	0.94	0.90	0.80	0.93	0.82	0.69	0.93	0.89	0.79
LDA	0.93	0.86	0.80	0.91	0.88	0.85	0.95	0.86	0.83	0.89	0.86	0.87

Abbreviations: AUC, area under the curve; LDA, linear discriminant analysis; NPV, negative predictive value; PPV, positive predictive value; SGD, stochastic gradient descent; SVC, support vector classification; SVM, support vector machine.

demonstrating the decrease of oleic acid, the product of linoleic acid, in the serum of lung cancer patients.[21] Coincidentally, metabolomic profile analyses in lipopolysaccharide (LPS)-induced acute lung injury rats also revealed a reduction in linoleic acid in both serum and lung tissue.[22] This may suggest some common signaling pathways in lung injury and lung tumor initiation. We also found that choline was dysregulated in NSCLC, which is one of the metabolites involved in cancer cell signaling and contributes to both cancer growth and programmed cell death.[23] In addition, we observed an increase in phenylalanine, an essential aromatic amino acid (AA), in the serum of patients with lung cancer. This finding was validated in an analysis of the AA profile in our other case–control study.[24] Moreover, prior work reported a higher concentration of phenylalanine in lung cancer tissues than paired paracarcinomous tissues,[17] and a correlation of this AA was also found between plasma and cancer tissues of lung cancer patients.[25]

The enrichment analysis revealed that the metabolism of phenylalanine, together with its derivates tyrosine and tryptophan were the most significantly enriched pathways. It was reported that the activity of phenylalanine hydroxylase (which converts phenylalanine to tyrosine) was dysfunctional in inflammatory or malignant disease states. This raises the possibility that circulating phenylalanine levels could be altered in patients with cancer due to the reduced action of this enzyme.[26,27] Moreover, a spatiotemporal intratumoral distribution of ring-$^{13}C_6$ labeled phenylalanine ($^{13}C_6$-Phe) and tyrosine ($^{13}C_6$-Tyr) in NSCLC showed higher abundances of $^{13}C_6$-Phe and $^{13}C_6$-Tyr in viable tumor regions in contrast with nonviable regions, and the incorporation of $^{13}C_6$-Phe in the tumor protein synthesis presented a different kinetics compared with the progressive incorporation in liver.[28] Metabolism of linoleic acid was one of the pronounced pathways in our analysis. Linoleic acid, the most abundant polyunsaturated fatty acid (PUFA) in diet, belongs to an ω-3 PUFA, which possess anti-inflammatory and antiallergic properties, and may thus be beneficial to lung health.[29] Several studies have reported that PUFAs and their derivatives inhibited the growth of human lung cancer cells, and triggered lung cancer cell apoptosis and autophagy.[30,31] Linoleic acid metabolism was also identified as one of the mainly altered pathways by acute lung injury.[22] Meanwhile, prospective studies have demonstrated an inverse association of PUFAs intake with lung cancer risk.[32] However, some other studies have reported contrary findings. Liu et al. found that serum linoleic acid and its oxidized metabolites were elevated in patients with LUAD.[33] A gene-expression profiling and multispectral imaging analysis have also identified a transitory increase in fatty acid metabolism in LUSC.[34] Therefore, the relationship between linoleic acid metabolism and lung cancer is still obscure until now. Of particular interest, we also found aberrant bile acid metabolism in NSCLC. Increasing evidence suggests that bile acid metabolism is critically important for maintaining a healthy gut microbiota, balanced lipid and carbohydrate metabolism, insulin sensitivity and innate immunity, and may play a role in the progression of gastrointestinal cancers like liver cancer and colorectal cancer.[35–37] Activation of G protein-coupled bile acid receptor 1 (GPBAR1, also

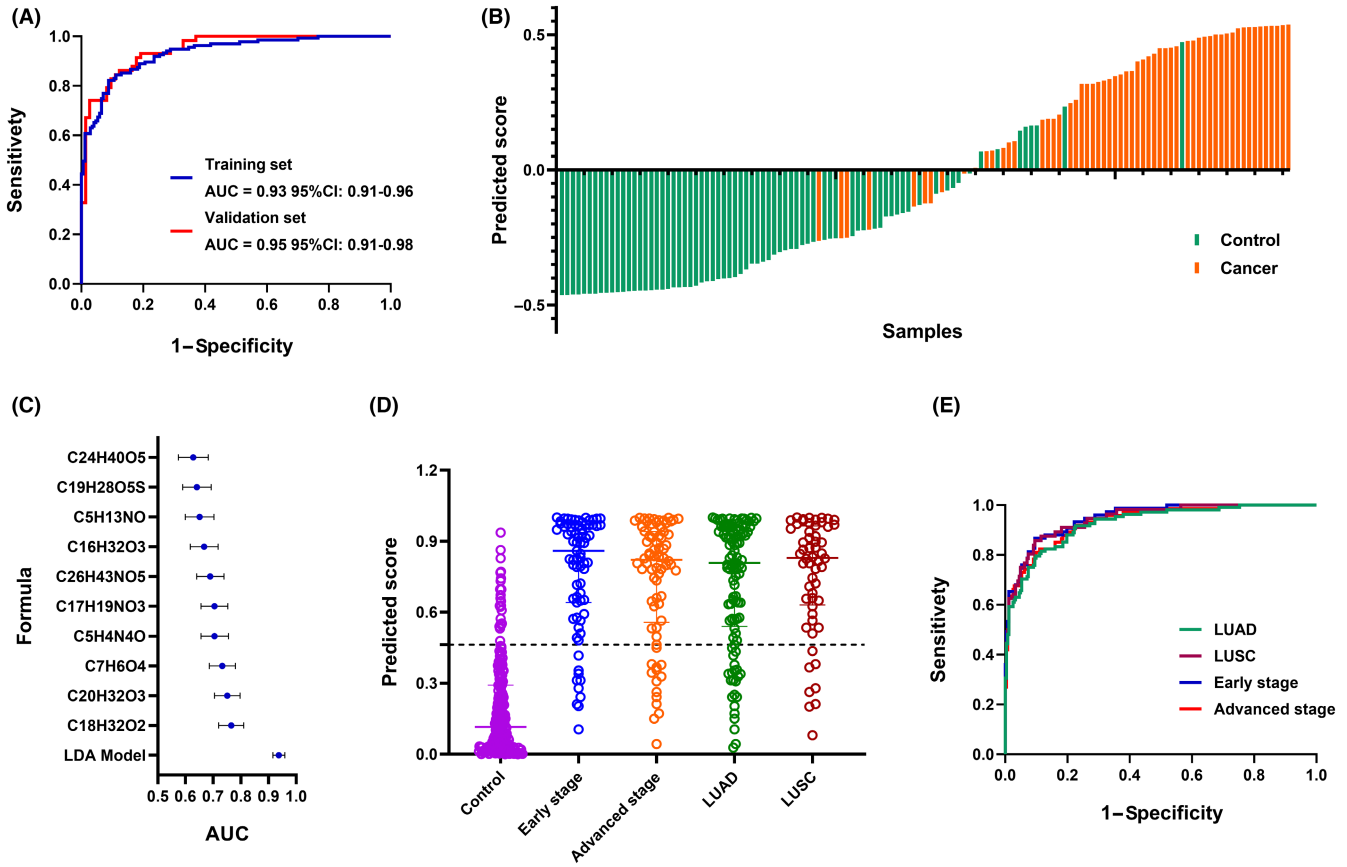

FIGURE 2 Diagnostic performance of a panel comprised of 10 metabolic biomarkers by a machine learning model using linear discriminant analysis (LDA). (A) Receiver operator characteristic (ROC) curve of the panel on the training set and validation set. (B) The distribution of predicted scores calculated by LDA model in cases and controls in the validation set. (C) Area under the curve (AUC) of single metabolite included in the model and the combined performance using all the subjects. (D) The predicted scores calculated by LDA model in different pathological types and clinical stages. (E) Classification performances of the panel for detection of different pathological types and clinical stages.

known as TGR5) induced by bile acids led to an increase in intracellular cyclic AMP (cAMP), thus triggering downstream signaling events that were associated with metabolic diseases and cancers.[38] Intriguingly, a recent metabolic landscape of lung cancer indicated that dysregulated bile acid metabolism facilitated the migration of LUAD, indicating that the dysfunction of bile acids might be a universal mechanism of carcinogenesis, not merely in gastrointestinal cancers.[39] In addition, an evaluation of tumor-derived metabolomic data suggested the involvement of bile acid biosynthesis in the therapeutic response of lung cancer.[40] Taken together, our findings highlight anomalous metabolic reprogramming during lung cancer initiation and progression. However, the biological mechanisms responsible for the perturbations of these pathways remain unclear in lung cancer and further investigations are required to explore these relationships.

Although some lung cancer screening approaches such as LDCT improved lung cancer detection in some degree, there is still a paucity of noninvasive and efficient biomarkers to refine the reliable and accurate diagnosis of lung cancer in an early stage. With the rapid advances in metabolomics technologies, nowadays metabolomic profiling becomes a promising tool to hunt for biomarkers available for early cancer detection. Surprisingly, the metabolic biomarkers screened from metabolomic profiling seem to exert good diagnostic efficiency in cancer diagnosis according to published literature. Using an unbiased metabolomics profiling approach, William and colleagues identified diacetyl spermine as a biomarker for NSCLC, with an AUC greater than 0.8 only using this single metabolite.[41] A quantitative formula LCAID v2.0 derived from nine lipids selected based on untargeted lipidomics achieved a striking accuracy of 94.96%, with a 100% specificity and a 92.93% sensitivity (AUC = 0.998) for lung cancer in an independent validation cohort.[42] In line with this evidence, our study revealed a panel of 10 metabolites with excellent differentiating capacity (AUC = 0.93 in the training set, and 0.95 in the validation set) for NSCLC classification based on machine learning. Of particular importance, the combination of 10 metabolites was demonstrated to have desirable performance not only for advanced NSCLC, but also for NSCLC within an early stage, suggesting the utility of this panel for lung cancer early detection.

It is still a pressing challenge to support lung cancer diagnosis in a noninvasive and convenient manner over recent years. Using an untargeted metabolic profiling-based screening, our study offers a

perspective on the identification of NSCLC metabolic alteration. The finding of the biomarkers might shed light on the clinical detection of lung cancer, especially for those cancers in an early stage, and additionally, the screening of high-risk crowds of cancer in Chinese population. However, limitations should be acknowledged. First, only a fraction of the altered metabolites was successfully identified, attributed to the difficulty of material identification generally encountered by the untargeted profiling approaches. The establishment of a standard library containing manifold chemicals was required. Second, although we have divided all the subjects into a training set and a validation set, an independent testing set was warranted to evaluate the external validity of our metabolic biomarkers in NSCLC recognition. Third, given that metabolism reprogramming is a common characteristic shared by solid tumors, even benign disorders, the application of our metabolic panel thus should be cautious. Further investigations are needed to determine whether it satisfied the differential diagnosis of NSCLC with other cancers, as well as with benign lung diseases such as pneumonia and chronic obstructive pulmonary disease (COPD).

5 | CONCLUSIONS

In summary, the serum metabolic pattern of NSCLC was massively changed from healthy conditions. The altered metabolites were enriched in pathways involving the metabolism of phenylalanine, linoleic acid, and bile acids. Moreover, we demonstrated a panel of 10 metabolic biomarkers that showed excellent discriminating capability for detection of NSCLC, in particular for early-stage lung cancer. Our work provides a paradigm of untargeted metabolisms combined with machine learning to characterize cancer metabolic alteration and refine disease diagnosis.

ACKNOWLEDGEMENT

We thank Tongji Hospital for providing clinical samples and the research platform for our study. This study was supported by National Key Research and Development Plan Program of China (No. 2016YFC1302702) and National Natural Science Foundation of China (No. 81572071, 81903394).

FUNDING INFORMATION

The work was funded by the National Key Research and Development Plan Program of China (Grant No. 2016YFC1302702) and the National Natural Science Foundation of China (Grant Nos. 81903394 and 81572071).

DISCLOSURE

The authors declare no conflict of interest.

ETHICAL APPROVAL

Approval of the research protocol by an Institutional Reviewer Board: This study was approved by the Institutional Review Committee of the Tongji Medical College, HUST.

Informed Consent: Written informed consent was obtained from all patients.

Registry and the Registration: N/A.

Animal Studies: N/A.

ORCID

Liming Cheng https://orcid.org/0000-0001-6444-5090

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.
2. Allemani C, Matsuda T, Di Carlo V, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. 2018;391(10125):1023-1075.
3. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med*. 2020;382(6):503-513.
4. Dudley JC, Diehn M. Detection and diagnostic utilization of cellular and cell-free tumor DNA. *Annu Rev Pathol*. 2021;16:199-222.
5. Cai G, Cai M, Feng Z, et al. A multilocus blood-based assay targeting circulating tumor DNA methylation enables early detection and early relapse prediction of colorectal cancer. *Gastroenterology*. 2021;161(6):2053-2056.e2.
6. Xiao Q, Zhang F, Xu L, et al. High-throughput proteomics and AI for cancer biomarker discovery. *Adv Drug Deliv Rev*. 2021;176:113844.
7. Martínez-Reyes I, Chandel NS. Cancer metabolism: looking forward. *Nat Rev Cancer*. 2021;21(10):669-680.
8. Wang Z, Cui B, Zhang F, et al. Development of a correlative strategy to discover colorectal tumor tissue derived metabolite biomarkers in plasma using untargeted metabolomics. *Anal Chem*. 2019;91(3):2401-2408.
9. Cai FF, Song YN, Lu YY, Zhang Y, Hu YY, Su SB. Analysis of plasma metabolic profile, characteristics and enzymes in the progression from chronic hepatitis B to hepatocellular carcinoma. *Aging (Albany NY)*. 2020;12(14):14949-14965.
10. Wei Y, Jasbi P, Shi X, et al. Early breast cancer detection using untargeted and targeted metabolomics. *J Proteome Res*. 2021;20(6):3124-3133.
11. Luo X, Liu J, Wang H, Lu H. Metabolomics identified new biomarkers for the precise diagnosis of pancreatic cancer and associated tissue metastasis. *Pharmacol Res*. 2020;156:104805.
12. Zhang L, Zheng J, Ahmed R, et al. A high-performing plasma metabolite panel for early-stage lung cancer detection. *Cancers (Basel)*. 2020;12(3):622.
13. Yang Z, Song Z, Chen Z, et al. Metabolic and lipidomic characterization of malignant pleural effusion in human lung cancer. *J Pharm Biomed Anal*. 2020;180:113069.
14. Kim HJ, Jang SH, Ryu JS, et al. The performance of a novel amino acid multivariate index for detecting lung cancer: a case control study in Korea. *Lung Cancer*. 2015;90(3):522-527.
15. Schmidt DR, Patel R, Kirsch DG, Lewis CA, Vander Heiden MG, Locasale JW. Metabolomics in cancer research and emerging applications in clinical oncology. *CA Cancer J Clin*. 2021;71(4):333-358.
16. Deja S, Porebska I, Kowal A, et al. Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. *J Pharm Biomed Anal*. 2014;100:369-380.
17. You L, Fan Y, Liu X, et al. Liquid chromatography-mass spectrometry-based tissue metabolic profiling reveals major metabolic pathway alterations and potential biomarkers of lung cancer. *J Proteome Res*. 2020;19(9):3750-3760.

18. Cífková E, Brumarová R, Ovčačíková M, et al. Lipidomic and metabolomic analysis reveals changes in biochemical pathways for non-small cell lung cancer tissues. *Biochim Biophys Acta Mol Cell Biol Lipids*. 2022;1867(2):159082.
19. Faubert B, Li KY, Cai L, et al. Lactate metabolism in human lung tumors. *Cell*. 2017;171(2):358-371.e9.
20. Liao ZX, Kempson IM, Hsieh CC, Tseng SJ, Yang PC. Potential therapeutics using tumor-secreted lactate in nonsmall cell lung cancer. *Drug Discov Today*. 2021;26(11):2508-2514.
21. Zheng Y, He Z, Kong Y, et al. Combined metabolomics with transcriptomics reveals important serum biomarkers correlated with lung cancer proliferation through a calcium signaling pathway. *J Proteome Res*. 2021;20(7):3444-3454.
22. Wang T, Lin S, Liu R, et al. Metabolomic profile perturbations of serum, lung, bronchoalveolar lavage fluid, spleen and feces in LPS-induced acute lung injury rats based on HPLC-ESI-QTOF-MS. *Anal Bioanal Chem*. 2020;412(5):1215-1234.
23. Ridgway ND. The role of phosphatidylcholine and choline metabolites to cell proliferation and survival. *Crit Rev Biochem Mol Biol*. 2013;48(1):20-38.
24. Liu K, Li J, Long T, et al. Changes in serum amino acid levels in non-small cell lung cancer: a case-control study in Chinese population. *PeerJ*. 2022;10:e13272.
25. Zhao Q, Cao Y, Wang Y, et al. Plasma and tissue free amino acid profiles and their concentration correlation in patients with lung cancer. *Asia Pac J Clin Nutr*. 2014;23(3):429-436.
26. Neurauter G, Grahmann AV, Klieber M, et al. Serum phenylalanine concentrations in patients with ovarian carcinoma correlate with concentrations of immune activation markers and of isoprostane-8. *Cancer Lett*. 2008;272(1):141-147.
27. Wiggins T, Kumar S, Markar SR, Antonowicz S, Hanna GB. Tyrosine, phenylalanine, and tryptophan in gastroesophageal malignancy: a systematic review. *Cancer Epidemiol Biomarkers Prev*. 2015;24(1):32-38.
28. Cao J, Balluff B, Arts M, et al. Mass spectrometry imaging of L-[ring-13 C 6]-labeled phenylalanine and tyrosine kinetics in non-small cell lung carcinoma. *Cancer Metab*. 2021;9(1):26.
29. Vega OM, Abkenari S, Tong Z, Tedman A, Huerta-Yepez S. Omega-3 polyunsaturated fatty acids and lung cancer: nutrition or pharmacology? *Nutr Cancer*. 2021;73(4):541-561.
30. Siena L, Cipollina C, Di Vincenzo S, et al. Electrophilic derivatives of omega-3 fatty acids counteract lung cancer cell growth. *Cancer Chemother Pharmacol*. 2018;81(4):705-716.
31. Zajdel A, Wilczok A, Tarkowski M. Toxic effects of n-3 polyunsaturated fatty acids in human lung A549 cells. *Toxicol In Vitro*. 2015;30(1 Pt B):486-491.
32. Luu HN, Cai H, Murff HJ, et al. A prospective study of dietary polyunsaturated fatty acids intake and lung cancer risk. *Int J Cancer*. 2018;143(9):2225-2237.
33. Liu J, Mazzone PJ, Cata JP, et al. Serum free fatty acid biomarkers of lung cancer. *Chest*. 2014;146(3):670-679.
34. Mascaux C, Angelova M, Vasaturo A, et al. Immune evasion before tumour invasion in early lung squamous carcinogenesis. *Nature*. 2019;571(7766):570-575.
35. Ma C, Han M, Heinrich B, et al. Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells. *Science*. 2018;360(6391):eaan5931.
36. Jia W, Xie G, Jia W. Bile acid-microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. *Nat Rev Gastroenterol Hepatol*. 2018;15(2):111-128.
37. Ocvirk S, O'Keefe SJD. Dietary fat, bile acid metabolism and colorectal cancer. *Semin Cancer Biol*. 2021;73:347-355.
38. Deutschmann K, Reich M, Klindt C, et al. Bile acid receptors in the biliary tree: TGR5 in physiology and disease. *Biochim Biophys Acta Mol Basis Dis*. 2018;1864(4 Pt B):1319-1325.
39. Nie M, Yao K, Zhu X, et al. Evolutionary metabolic landscape from preneoplasia to invasive lung adenocarcinoma. *Nat Commun*. 2021;12(1):6479.
40. Miller HA, Yin X, Smith SA, et al. Evaluation of disease staging and chemotherapeutic response in non-small cell lung cancer from patient tumor-derived metabolomic data. *Lung Cancer*. 2021;156:20-30.
41. Wikoff WR, Hanash S, DeFelice B, et al. Diacetylspermine is a novel Prediagnostic serum biomarker for non-small-cell lung cancer and has additive performance with pro-surfactant protein B. *J Clin Oncol*. 2015;33(33):3880-3886.
42. Wang G, Qiu M, Xing X, et al. Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci Transl Med*. 2022;14(630):eabk2756.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Li J, Liu K, Ji Z, et al. Serum untargeted metabolomics reveal metabolic alteration of non-small cell lung cancer and refine disease detection. *Cancer Sci*. 2023;114:680-689. doi: 10.1111/cas.15629