Phenylallylamylphenylalanine as a Diagnostic Biomarker for Lung Cancer and Tuberculosis.

Siyu Chen  
Tianjin University of Traditional Chinese Medicine

Chunyan Li  
Tianjin University of Traditional Chinese Medicine

Zhonghua Qin  
Tianjin Haihe Hospital

Lili Song  
Tianjin University of Traditional Chinese Medicine

Shiyuan Zhang  
First Teaching Hospital of Tianjin University of Traditional Chinese Medicine

Chongxiang Sun  
First Teaching Hospital of Tianjin University of Traditional Chinese Medicine

Pengwei Zhuang  
Tianjin University of Traditional Chinese Medicine

Yuming Wang  
Tianjin University of Traditional Chinese Medicine

Bin Yang  
Tianjin University of Traditional Chinese Medicine

Yanjun Zhang  
Tianjin University of Traditional Chinese Medicine

Yubo Li  
Tianjin University of Traditional Chinese Medicine

Research

Keywords: metabonomic analysis, lung cancer, tuberculosis, differential diagnosis

DOI: https://doi.org/10.21203/rs.3.rs-775519/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License
Abstract

Background: Worldwide, lung cancer has the highest mortality rate, and pulmonary tuberculosis has a high incidence in China, and both may be misdiagnosed frequently because of similar clinical presentation and atypical imaging findings. Diagnostic biomarkers to distinguish between lung cancer and other pulmonary diseases can be detected by metabolomics to avoid non-essential treatment.

Methods: This cohort study employed non-targeted and targeted metabolomic analysis in participants enrolled from three independent centers. Multivariate statistics, variable importance in the projection parameter, receiver operating characteristics (ROC) curve were used to build potential key diagnostic biomarkers model of lung cancer and these were subsequently analyzed using targeted metabolomics in test set. Quantitative analysis of differences in biomarker levels was conducted, and a support vector machine (SVM) classifier was used to identify the prediction rate of diagnostic biomarker model.

Results: Phenylalanylphenylalanine showed opposite trends in lung cancer and tuberculosis. The area under the curve 0.8887 (95% CI 0.8064–0.9710, p<0.001, sensitivity 85.45%, specificity 84%), 0.8149 (95% CI 0.7419–0.8878, p<0.001, the sensitivity was 73.26%, the specificity was 78.43%) and SVM results (prediction rate 77.94%) showed the feasibility of using phenylalanylphenylalanine as a diagnostic marker for the differential diagnosis of lung cancer and tuberculosis.

Conclusions: Changes in the levels of phenylalanylphenylalanine facilitate differential diagnosis between lung cancer and tuberculosis, thereby potentially reducing the damage caused by misdiagnosis in the clinical setting, and enabling early treatment of lung cancer patients.

Trial registration: This study is registered in the China Clinical Trial Registration Center (registration number ChiCTR2000040666, Registered 07 December 2020, http://www.chictr.org.cn/index.aspx)

Background

Lung cancer (LC) is a malignant respiratory system tumor and has been the commonest cause of cancer mortality worldwide for nearly half a century [1]. According to the International Agency for Research on Cancer, LC has the highest incidence rate in the world (1.8 million, 13% of the total), and in terms of mortality, LC once again ranks first (1.6 million, 19.4% of the total) [2]. Statistics show that the 5-year survival rate of lung cancer is only 18% [3]. In the past few decades, many scientists and medical researchers have undertaken great efforts to control the occurrence of lung cancer [4]. If LC is detected at an early stage, the 5-year survival rate of lung cancer can be increased by 70% [5]. Due to the occult early onset of LC, the clinical symptoms lack specificity; therefore, it is difficult to distinguish LC from other lung diseases and most patients present with advanced-stage LC at diagnosis, and this adversely affects the prognosis and treatment of patients [6]. LC and atypical tuberculosis (TB) have similar clinical symptoms and examination findings, and often evince the phenomenon of "different diseases and same shadows" in imaging investigations as some TB lesions resemble cancer. In some non-tumor hospital in China, LC is often misdiagnosed, further increasing the difficulty for an accurate diagnosis by clinicians
[7–11]. 18F-FDG-PET/CT for molecular/atomic imaging is an important tool for the detection, identification, and staging of LC and has broad clinical application, but is expensive and difficult to widely popularize. Furthermore, in areas where tuberculosis is endemic, the false positive rate is high and the specificity is low [12]. The gold standard for the clinical diagnosis of LC is histopathology of biopsy specimens but this method is invasive which poses a risk to middle-aged and elderly people. Therefore, safe methods for early diagnosis of LC are needed for clinical application [13].

For a more comprehensive understanding of LC pathogenesis, to improve diagnostic accuracy, and to reduce LC mortality rates, researchers have studied molecular biomarkers in genomics, epigenomics, proteomics, and metabolomics to facilitate the early detection of LC [14–17]. In recent years, the identification of efficient tumor markers for clinical diagnosis has become a research hotspot, and tumor markers that have been found to be effective and have wide clinical application include: carcinoembryonic antigen, cytokeratin-19-fragment, and neuron-specific enolase [18], however, the lack of accuracy of these markers limits their usefulness for the early diagnosis of LC. Metabolites can reflect the human physiological functions and pathological characteristics in great detail. Metabolomics is a powerful tool to analyze the differences in metabolites between the healthy and diseased populations. The rise of bioinformatics has resulted in the rapid progress of metabolomics in the past decade. Metabolomics can enable qualitative or quantitative analysis of metabolites and facilitate the identification of unknown metabolites [19, 20]. The use of metabolomics technology to find metabolic biomarkers for use in supplementary metabolomics analysis is an effective method to detect disease onset, and can possibly provide new options for clinical diagnosis. Furthermore, metabolomics can monitor response changes of key biomolecules in patients, and noninvasive methods using metabolic markers for clinical diagnosis have been widely used for the identification of tumor markers for pancreatic [21] and ovarian [22] cancer, which has important implications for establishing early diagnostic methods for LC. This study aimed to identify metabolic markers that differ significantly between LC and TB patients through non-targeted metabolonomic analysis and to undertake subsequent quantitative targeted metabolonomic analysis to find metabolic markers that can be applied in actual clinical practice, to facilitate a more accurate clinical diagnosis of LC, prevent clinical misdiagnosis events, and enable timely and effective treatment of LC patients.

**Methods**

**Study Design**

In training set and identification set, we prospectively enrolled TB and LC patients and non-cancer patients as participants for the serum-based non-targeted metabolomics analysis to detect specific markers of both diseases and the identified markers were screened through a binary logistic regression model. In test set, targeted metabolomics technology was used to analyze differences among TB, LC, and pneumonia patients and in non-cancer controls. The SVM model evaluated the predicting rate of potential biomarkers in LC and TB, and provide data for clinical research.
Participants

The study participants were patients who were treated at three centers: LC patients (n=262) from the Second Hospital of Tianjin Medical University; TB patients from Tianjin Haihe Hospital (n=182); and pneumonia patients from the First Affiliated Hospital of the Tianjin University of Chinese Medicine (n=30). The non-cancer control group (n=218) comprised patients with noncancerous diseases who were treated at the Second Hospital of Tianjin Medical University during the same period and were confirmed to have no malignant disease based on a review of the hospital’s medical records from the hospital’s medical database. The prespecified blood sampling and sample preparation were undertaken by well-trained researchers. The subject has passed the review of the ethics committee of the Second Hospital of Tianjin Medical University. The age and clinical manifestations of each participant were ascertained, and the basic information of the matched participants is presented in Table 1. This study is registered in the China Clinical Trial Registration Center (registration number ChiCTR2000040666, Registered 07 December 2020, http://www.chictr.org.cn/index.aspx), and the registration unit is the Second Hospital of Tianjin Medical University.

Table 1 Basic information of the participants

|                      | Training set | Identification set |
|----------------------|--------------|--------------------|
|                      | TB | LC | NC | TB | LC | NC |
| Total                | 30 | 55 | 35 | 63 | 88 | 63 |
| Female               | 15 | 24 | 18 | 41 | 59 | 43 |
| Male                 | 15 | 31 | 17 | 22 | 29 | 20 |
| Age                  | 59.27±8.33 | 58.58±5.61 | 59.83±5.62 | 62.21±9.29 | 63.28±7.60 | 62.68±9.31 |
| TCT(+/−/NA)          | 15/14/1 | /  | /  | 27/29/7 | /  | /  |
| AFB(+/−/NA)          | 15/15/0 | /  | /  | 20/37/6 | /  | /  |
| Stage                | /  | /  | /  | /  | /  | /  |
| I                    | /  | 0  | /  | /  | 2  | /  |
| II                   | /  | 8  | /  | /  | 28 | /  |
| III                  | /  | 13 | /  | /  | 30 | /  |
| IV                   | /  | 8  | /  | /  | 28 | /  |

Abbreviations: NC, non-cancer control; TB, tuberculosis; LC, lung cancer; TCT, tuberculin culture test; AFB, AFB sputum smear.

Inclusion criteria: Patients who were diagnosed based on diagnostic criteria for TB and LC, with pathological confirmation of LC in biopsy specimens or a clinician-confirmed LC based on radiological
and clinical findings. According to the 7th edition of the TNM classification, all LC participants had primary lung cancer. The TB patients were diagnosed based on positive results on sputum examination (sputum smear or culture) and evidence of pulmonary TB on chest X-ray or computed tomography scanning. Patients in the age range of 18 to 80 years, regardless of sex; unimpaired consciousness level.

**Exclusion criteria:** 1) other metabolic diseases, hematological diseases, and cancer; 2) severe infection; and 3) severe cardiovascular, cerebrovascular, liver, or kidney dysfunction.

This study was conducted in accordance with the Declaration of Helsinki and the ethical rules of good clinical practice, and has been approved (approval number KY2020K089) by the ethics committee of the Second Hospital of Tianjin Medical University. The overall sample was randomly subdivided into training, identification, and test sets. Non-targeted metabonomics was used in training and identification set, targeted metabonomics was used in test set(Figure 1).

**Non-targeted metabonomic**

**Sample Preparation**

The samples that were frozen and stored at −80°C were completely thawed in the refrigerator at 4°C, and 80 μL serum was mixed with acetonitrile (chromatography-grade acetonitrile, Oceanpak, Sweden) at a 1:3 volume ratio. After swirling (LP Vortex Mixer, Thermo Fisher Scientific (China) Co., Ltd.) for 1 min, the samples were ultrasonicated (JP-060S ultrasonic cleaner, Jiemeng Cleaning Equipment Co., Ltd.) in an ice-water bath for 10 min and centrifuged (3-18K high-speed centrifuge, Germany Sigma) at 4°C for 15 min at 13000 rpm; then, 200 μL supernatant was obtained for metabonomic analysis. The supernatants from the training and identification set were tested in ultra-high-performance liquid chromatography-Quadrupole Time-Of-Flight Mass Spectrometry (UPLC/Q-TOF-MS) analysis in positive mode. Each serum sample was pipetted into a centrifuge tube, and then mixed and vortexed for 1 min to prepare a quality control (QC) sample. The QC sample contained the biological information of all samples and can reflect the overall sample status [23] and was used for methodological investigation with the same pre-processing method as that used for the sample.

**Mass Spectrometric Analysis**

Using electrospray ionization (ESI source), mass spectrometric analysis was performed in positive ionization mode with the following conditions: capillary voltage 2.0 kV, ionization source temperature 100°C, dry gas flow rate 10 mL/min, esolavation flow rate 600 L/D, desolvation temperature 450°C, cone air flow rate 50 L/D, and quadrupole scanning range m/z 50–1000.

**Chromatographic Analysis**

In this experiment, Waters ACQUITY UPLC (Waters Corporation, USA) was used for metabolomics research under the following conditions: chromatographic column: UPLC BEH C18 (2.1×100 mm, 1.7 μm, Waters Corporation, Milford, USA), column temperature: 45°C; flow rate: 0.3 mL/min; injection volume: 5
μL; and mobile phase composition: A: 0.1% formic acid (chromatography-grade formic acid (ROE, USA)) in water and B: 0.1% formic acid in acetonitrile. The elution gradient was 0.0–0.5 min, 1% B; 0.5–2.0 min, 1%–50% B; 2–9 min, 50%–99% B; 9–10 min, 99% B; 10.0–10.5 min, 99%–1% B; and 10.5–12.0 min, 1% B.

**Procedures of Methodological Investigation**

*Instrument precision test:* We took the same QC sample and made 6 consecutive injections. The data were exported as the peak area, after the missing value was filled, and the relative standard deviation (RSD) value of each ion feature was calculated; the features with <30% accounted for >70%.

Method precision test: Six QC samples were prepared in parallel and continuously injected for analysis. The data were exported as peak area, after the missing value was filled, the RSD value of each ion feature was calculated, and the features with RSD <30% accounted for >70%.

Sample stability test: We took the same QC sample and injected samples at six time points during the whole injection process. The data were exported as peak area, after the missing value was filled, the RSD value of each ion feature was calculated, and the features with RSD <30% accounted for >70%.

**Statistical Analysis**

UPLC-Q-TOF/MS technology was used to analyze the metabolomic profile of clinical serum samples in the NC, TB, and LC groups. Data collected by the instrument were passed through the Masslynx (version 4.1) data processing system (software parameters: mass error 0.01 da; retention time error 0.5 min), the intensity of each ion was normalized to the total number of ions, and the data formed included retention time, m/z value, and peak area. These data were imported into the SIMCA-P version 14.1 (Umetrics, Sweden) for multivariate statistical analysis. Unsupervised principal component analysis (PCA) and supervised partial least squares discriminant analysis (PLS-DA) models were established to identify potential discriminant variables[24]. The PCA model was used to eliminate outlier samples. According to a variable importance in the projection (VIP) parameter >1 of the metabolic ion in the PLS-DA model, the T-test was used to screen out substances with $p<0.05$ to indicate the least-different molecular metabolites. We used the m/z value to search through HMDB (http://www.hmdb.ca/), and preliminarily screened out different small molecular metabolites based on fragment information.

**Targeted Metabolomic Analysis**

*Sample preparation:* Samples frozen and stored at −80°C were thawed completely in the refrigerator at 4°C; then, 5 μL samples were taken into an EP tube, 995 μL methanol was added, vortexed for 1 min, allowed to stand for 5 min, centrifuged at 13,000 rpm for 10 min, and the supernatant was obtained for injection analysis. An appropriate amount of methanol was added into the weighed mixture to dilute 250, 100, 50, 25, 10, 5, 2.5, 1, and 0.5 ng/mL series solution for use as the standard solution.
Conditions of mass spectrometry analysis: The source voltage was ES+3.00 KV, source temperature was 400°C, gas-flow rate was 700 L/h, and cone was 50 L/h. The ion pair information of phenylalanylphenylalanine is 313.1/119.9, the cone is 30V, the collision is 18V.

Conditions of chromatographic analysis: In this experiment, UPLC (I-class)-MS (xexo TQD) was used for quantitative analysis. The chromatographic column was Waters ACQUITY UPLC BEH C18; column and sample-chamber temperatures: 35°C and 4°C, respectively. The mobile phase comprised A: 0.1% formic acid in water; B: 1% formic acid water in acetonitrile; the flow rate was 0.3 mL/min, and the injection volume was 1 μL. The specific elution gradient was 0–1 min, 5% B; 1.0–2.3 min, 5%–15% B; 2.3–3.0 min, 15%–100% B; 3.0–4.5 min, 100% B; 4.5–5.0 min, 100%–5% B; and 5–7 min, 5% B.

Methodology investigation: The standard curve was detected every day to determine the linearity and the minimum limit of quantitation.

Results

Non-targeted Metabonomics Methodology

Instrument precision test: The RSD <30% accounted for 86.27%, >80%. Method precision test: The RSD <30% accounted for 80.88%, >80%. Sample stability test: The RSD <30% accounted for 73.88%, >70%. The method verification results prove that the instrument precision, method repeatability, and sample stability all met the requirements of metabolomic research. Serum samples are shown in the positive-ion mode base peak intensity diagram (Figure 2).

Metabolic Global Analysis

First, we used Origin 2021 to conduct a global metabolomic analysis through the volcano map, which is conducive to describing the overall situation of the substance. Using log₂ fold-change as the abscissa and p-value (-log₁₀) as the ordinate to create a volcano map, we set the threshold to 1.2, and each point represents a substance. The larger the ordinate and the smaller the p-value, the more significant was the difference. The closer the abscissa was to both sides, the greater was the upregulation or downregulation of the substance (Figure 3).

Multivariate Statistical Analysis

The serum sample information was collected by UPLC-Q-TOF/MS, and the multivariate pattern recognition method was used to process the data, and SIMCA-P was used to perform multivariate statistical analysis on the complex data obtained from the experiment through dimensionality reduction. First, we perform unsupervised principal component analysis (PCA) model analysis on the data in the positive-ion mode for the LC, NC and TB groups. PCA was used to determine the separation of samples and to eliminate outliers and the differential metabolites between groups are determined by partial Least Squares Discrimination Analysis (PLS-DA). Both the LC and TB groups and the control group showed
obvious classification and aggregation on the scatter diagram (Figure 4), indicating that the metabolic pattern of the disease group is quite different from that of the control group. The effect evaluation standard of the PLS-DA model is established based on the $R^2_Y$ and $Q^2$. The greater the significant between-group difference, the closer the $R^2_Y$ value and $Q^2$ value is to 1. In this study model, the $R^2_X=0.163$, $R^2_Y=0.772$, and $Q^2=0.567$ were noted for the LC and NC groups; $R^2_X=0.157$ and $R^2_Y=0.916$, and $Q^2=0.692$ were noted for the TB and NC groups. The abovementioned results show that the prediction model established in this experiment has good fit and the prediction results are stable and reliable. In this study, in the NC-TB or NC-LC, the PLS-DA model had no over-fitting phenomenon and was therefore reliable and predictive (Figure 4).

Through the search of training set and the verification of identification set, the validation criteria are: the markers can be reproduced in the identification set and consistent with the change trend of the training set, we finally determined 30 biomarkers—the results are shown in Table S1. Among them, L-phenylalanine, phenylalanylphenylalanine, and 1-methylinosine were compared with the standard product (Phenylalanylphenylalanine (98%) and 1-methylcreatinine (≥99%) were purchased from Ron reagent (Shanghai Yien Chemical Technology Co., Ltd.). L-phenylalanine (≥98%) was purchased from Shanghai Yuanye Biotechnology Co., Ltd.). The MetPA (https://www.metaboanaly-st.ca/) database provides a free online metabolomic metabolic pathway analysis tool, which is often used for metabolic pathway analysis related to differential metabolic biomarkers [25]. The results of the MetPA analysis are shown in Figure 5, and mainly involves porphyrin and chlorophyll metabolism; glycerophospholipid metabolism; phenylalanine, tyrosine, and tryptophan biosynthesis; phenylalanine metabolism, etc., based on which we can draw a network diagram (Figure 6) to clearly visualize the interactive relationship among various metabolites.

**Biomarker Analysis**

*Cluster analysis*

To observe the relative changes of biomarker levels in different groups more intuitively and concretely, we used hierarchical cluster analysis—heat map to analyze the key biomarkers (eight differential markers with common significant changes in LC and TB groups). In Figure 7, the horizontal axis represents sample information, and the vertical axis represents the biomarkers. The color block reflects the value of the variable: the higher the content, the darker the color block. The closer bifurcation of the left vertical axis indicates that the similarity of these substances is higher—that is, these metabolites may originate from the same substance. From Figure 7, it is apparent that compared with the non-cancer controls the contents of these metabolites significantly differed between the LC or TB groups. Among them, L-phenylalanine and phenylalanylphenylalanine showed a significant upward and downward trend in the LC and TB groups, respectively. Thus, the multiple of change is significant.

In the identification set, both L-phenylalanine and phenylalanylphenylalanine showed similar change trends to those in the training set. The change multiple of L-phenylalanine in the identification set slightly
differed from that in the training set, which may be related to the increased sample size. But from the clustering, the great similarity of L-phenylalanine and phenylalanylphenylalanine is evident. Therefore, in the next study, we plan to use L-phenylalanine and phenylalanylphenylalanine as potential biomarkers to analyze and then determine the final biomarker.

**Evaluation of Clinical Performance**

To explore whether these markers have diagnostic significance, we used graphpad 8.3.0 and conducted receiver operating characteristics (ROC) curve analysis to verify the diagnostic ability of the abovementioned screening substances, establish an effective diagnosis model. Sensitivity and 1 – Specificity were the ordinate and abscissa, respectively, and the relationship between the sensitivity and specificity of the markers is visually observable as the diagnostic efficacy of markers in the area under the curve (AUC) corresponding to each marker. The metabolite information of the two biomarkers screened in LC and TB were used to establish the binary logistic regression model, and the AUC value of each biomarker was obtained. Combining the results of the training and identification sets(Figure 8), the AUC of phenylalanylphenylalanine is 0.8887 (95% CI 0.8064–0.9710, p<0.001, the sensitivity is 85.45%, the specificity is 84%) and 0.8149 (95% CI 0.7419–0.8878, p<0.001, the sensitivity was 73.26%, the specificity was 78.43%) All>0.8, phenylalanylphenylalanine has greater diagnostic significance. The AUC of L-phenylalanine was 0.8615 (95% CI 0.7771–0.9459, p<0.001, the sensitivity was 81.48%, the specificity was 88%) and 0.5889 (95% CI 0.4927–0.6851, p=0.50, the sensitivity was 19.77%, the specificity was 98.04%). Based on the above analysis, we selected phenylalanylphenylalanine as diagnostic model, then as a potential diagnostic biomarker for the subsequent analysis.

**Targeted metabolomics analysis**

In order to analyze the level of markers in blood and the trend of changes among groups more accurately, 358 patients were analyzed using targeted metabolomics analysis with higher specificity and sensitivity. The test set included 89 TB, 119 LC, 30 pneumonia (P), and 120 non-cancer controls. To test the specificity of the marker in the differential diagnosis of LC, we added 30 pneumonia patients.

The results of standard and serum samples are shown in Figure 9. The lowest limit of quantification (LLOQ) of phenylalanylphenylalanine is 0.5 ng/mL. The results of the standard curve analysis are shown in Table 2. The marker levels in each group are shown in Table 3 and Figure 10.

**Table 2** Standard curve, linear range, and lowest limit of quantification analysis

| Name                              | Linear range (ng mL\(^{-1}\)) | Regression equation | \(R^2\) | Retention time (min) | LOQ (ng mL\(^{-1}\)) |
|-----------------------------------|-------------------------------|---------------------|---------|----------------------|----------------------|
| Phenylalanylphenylalanine         | 0.5-250                       | \(y = 146.02x + 63.40\) | 0.9986  | 3.48                 | 0.5                  |
**Table 3.** Change trend and multiples of phenylalanylphenylalanine in each group

| Name                     | TB/NC | LC/NC | P/NC |
|--------------------------|-------|-------|------|
| Phenylalanylphenylalanine| 0.8*  | 2.92**| 0.69*|

Note: **p<0.01, *p<0.05.

**Support Vector Machine Model Prediction**

A support vector machine prediction model was established by Matlab R2013a to judge the feasibility of the prediction. The samples of test set were divided into two groups: Group 0 (TB) and Group 1 (LC). The blood sample content in each group of samples was used as the input variable of the support vector machine to verify the accuracy and specificity of the markers. With two-thirds of the samples as the model's training set, the remaining one-third was used as the test set to determine the model's predictive accuracy. The support vector machine model parameters in the cross-validation method of phenylalanylphenylalanine are shown in Figure 11. The accuracy of the model can reach 83.57%, which indicates that the model is stable and reliable. The results show that the prediction rate reaches 77.94%, which indicates that the biomarker diagnostic model has high predictive accuracy, good predictive ability, stability.

**Discussion**

This study describes a comprehensive metabolomics assessment of LC and TB patients from three independent centers who were divided into the training, identification, and test sets. A non-targeted metabolomics evaluation based on UPLC-Q-TOF/MS technology initially explored the small-molecule metabolites in samples from LC and TB patients, and found significant differences in the metabolic phenotypic pattern between LC and TB patients. Through the search of training set and the verification of identification set, eight metabolic markers with significant differences were screened out. Among them, L-phenylalanine and phenylalanylphenylalanine showed opposite trends in LC (upward) and TB (downward), respectively. Therefore, these two may be ideal biomarkers for the identification of LC and TB. In the training set and identification set, the AUCs of phenylalanylphenylalanine were 0.8887 and 0.8149 and that of L-phenylalanine were 0.8615 and 0.5889, respectively. Through the binary logistic regression model of the identification set, we selected phenylalanylphenylalanine as a potential diagnostic marker. The test set showed a downward trend for phenylalanylphenylalanine in TB patients (change multiple 0.8) and an upward trend in LC patients (change multiple 2.92). Pneumonia patients had the same trend as TB patients, but the opposite trend to that of LC patients (change multiple 0.69, indicating good specificity for LC). Support vector machine (SVM) model analysis showed that the model prediction rate of TB and LC was 77.94%, whereby phenylalanylphenylalanine provides excellent predictive value for the differential diagnosis between LC and pulmonary TB.
At present, the diagnosis of malignant tumors mainly relies on invasive biopsy. With the continuous in-depth development of modern bioinformatics technology, “noninvasive blood tests” have emerged as the need of the hour. Researchers have successively explored and discovered various tumor markers. Metabolic markers are specifically expressed in the blood of cancer patients, often before the onset of clinical symptoms and imaging manifestations [26]. However, there is no specific marker for LC at present, and the development of noninvasive, accurate biomarkers to determine the LC risk can facilitate early diagnosis and prolong survival.

Amino acids, the structural unit of protein synthesis in cells, provide energy for the growth of organic cells while maintaining normal body metabolism, and have always been considered as potential biomarkers of malignant diseases. Although glutamine is a major nutrient for tumor cell growth, studies have found that the ability of glutamine to supplement the intermediates of the tricarboxylic acid cycle is mainly used in the biosynthesis of amino acids. For example, aspartic acid and serine can not only contribute a large number of carbon atoms to the formation of new cells, but also constitute proteins. Therefore, amino acids are considered the most important source of nutrition for tumor cells [27]. Cancer development depends on amino acids. New cancer cells require a large amount of biomolecules, such as proteins, nucleic acids, lipids, and important cofactors to maintain the redox state of the cells. The dietary elimination of amino acids causes a drop in the level of certain amino acids, with a resultant slowing of cancer development as tumor cells cannot obtain sufficient energy and have limited speed of growth [28].

Phenylalanine metabolism is one of the metabolic pathways with abnormal levels in various cancers, and changes in the blood levels of phenylalanine have attracted attention. Some clinical studies have shown that, compared with healthy controls, patients with non-small cell lung cancer have higher levels of phenylalanine and ornithine [29]. A study found that phenylalanine concentrations decreased in patients with oral cancer, whereas another study found an increasing trend [30,31]. Phenylalanine hydroxylase catalyzes the conversion of phenylalanine to tyrosine, and the active function of the enzyme is dysfunctional in inflammatory states or in malignant diseases [32,33]. A search of the Kyoto Encyclopedia of Genes and Genomes database showed that phenylalanine may be transformed into phenylpyruvate by aspartate transaminase, histidine phosphate transaminase, phenylalanine (histidine) transaminase, tyrosine transaminase, aromatic amino acid transaminase, phenylalanine dehydrogenase, and other enzymes; phenylpyruvate is closely related to the tricarboxylic acid cycle [34]. As a glycogenic amino acid, phenylalanine can produce glucose in the tricarboxylic acid cycle and via gluconeogenesis, which can be used as an energy source during the rapid proliferation of cancer cells. We found that L-phenylalanine and phenylalanylphenylalanine levels increased in LC patients; therefore, we speculated that these amino acids might be related to the tricarboxylic acid cycle and may provide the energy needed for the growth of cancer cells. The dipeptide is the short-lived intermediate produced when the protein is hydrolyzed through an amino acid degradation pathway, and it is the product of dehydration and condensation of amino acids, which has strong physiological activity. Dipeptide metabolites are products of incomplete breakdown in protein catabolism that may be related to the removal of nutrients as well as
intracellular and extracellular protein catabolism; thus, they may contribute to tumor growth [35]. Glutamylleucine, a dipeptide, plays a role in the development of the metabolic syndrome an in colorectal cancer [36] and reflects oxidative stress and the inflammatory state of the metabolic syndrome. Phenylalanine has been found in thyroid cancer [37], but showed a downward trend, and it is also found in this study, however, there is an upward trend in the lung cancer group, which also proves the specificity of the markers we found for lung cancer. In this study, serum phenylalanylphenylalanine levels of LC patients increased significantly, but decreased significantly in the serum of TB patients, possibly due to the inflammatory reaction of LC and resultant stress in the body. However, the role of phenylalanylphenylalanine in LC and its related mechanisms need to be explored.

An advantage of this study is the large number of samples. The qualitative and quantitative analyses of markers were conducted with non-targeted and targeted metabonomics, and the biomarkers were found to be stable and reliable. However, the specificity and sensitivity of these biomarkers need to be determined in long-term follow-up experiments, and more large-scale, multicenter, and even multi-region sample studies are needed before their clinical application.

Conclusions

We found and verified metabolic markers phenylalanylphenylalanine based on UPLC-Q/TOF-MS in non-targeted metabolomics to distinguish LC and TB patients, and was selected as the optimal diagnostic biomarker by binary logic regression model. It showed opposite trends in LC and TB patients and. In the test set, the results of quantitative analysis of targeted metabolism showed results that were consistent with the previous experimental stage, and the change multiple indicated a large and significant difference. The SVM algorithm using machine learning showed that phenylalanylphenylalanine facilitates differential diagnosis between LC and TB, the classification prediction rate is fairly good, and reducing the irreversible injury caused by misdiagnosis in the clinical setting, and enabling early treatment of LC patients.

Abbreviations

LC, lung cancer; TB, tuberculosis; NC, non-cancer patients; PCA, principal component analysis; PLS-DA, partial least squares discrimination analysis; RSD, relative standard deviation; AUC, area under the curve; QC, quality control; UPLC/Q-TOF-MS, ultra-high-performance liquid chromatography-Quadrupole Time-Of-Flight Mass Spectrometry; SVM, support vector machine; VIP, variable importance in the projection

Declarations

Ethics approval and consent to participate
This study was conducted in accordance with the Declaration of Helsinki and the ethical rules of good clinical practice, and obtained relevant approval by the Ethics Committee of the Second Hospital of Tianjin Medical University. Ethics number KY2020K089.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Tianjin Science and Technology Plan Science and Technology Commission-Metabolomics and Chinese Medicine New Drug Team (KZ1901).

Authors' contributions

Conception and design: SC, CL, YZ, YL. Financial support: YL. Collection and process of samples: SC, CL, ZQ, SZ, CS, PZ, YW. The detection of samples: SC, LS, BY. Writing the paper and interpretate the results: SC, CL. All of authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424. doi:10.3322/caac.21492.
2. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-86. doi:10.1002/ijc.29210.
3. Tanoue LT, Tanner NT, Gould MK, Silvestri GA. Lung cancer screening. Am J Respir Crit Care Med. 2015;191:19–33. doi:10.1164/rccm.201410-1777CI.
4. Balata H, Fong KM, Hendriks LE, Lam S, Ostroff JS, Peled N, et al. Prevention and early detection for NSCLC: advances in thoracic oncology 2018. J Thorac Oncol. 2019;14:1513–27.
5. Wardwell NR, Massion PP. Novel strategies for the early detection and prevention of lung cancer. Semin Oncol. 2005;32:259–68. doi:10.1053/j.seminoncol.2005.02.009.

6. Granville CA, Dennis PA. An overview of lung cancer genomics and proteomics. Am J Respir Cell Mol Biol. 2005;32:169–76. doi:10.1165/rcmb.F290.

7. Bai M, Guo Y, Xie Dilai, Qin J. CT. pathological manifestations and misdiagnosis analysis of pulmonary tuberculosis misdiagnosed as lung cancer. Chin J CT MRI. 2012;10:52–4. in Chinese.

8. Xu S, Cheng L, Yang Xuanqin, Xin L. Shangfei fan. Analysis of CT findings of 33 cases of atypical pulmonary tuberculosis. Chin J Tuberc Prev. 2013;35:930–3. in Chinese.

9. Wang Z. Mingquan, Li. CT misdiagnosis analysis of 33 cases of atypical lung cancer. Chin J Pract Diagn Treat. 2012;26:378–9. in Chinese.

10. Gu G, Miao J, Zhang H. Retrospective analysis of CT findings of 48 cases of mass pulmonary tuberculosis misdiagnosed as lung cancer. Imaging research medical application. 2018;2:59–60. in Chinese.

11. Zhang X, Wang Q, Wang Z, Zhao H, Zou Q, Jia C, et al. Clinical analysis of 19 cases of pulmonary tuberculosis misdiagnosed as peripheral lung cancer by PET-CT. Chin J Exp Diagn. 2018;22:1762–3. in Chinese.

12. Niyonkuru A, Chen X, Bakari KH, Wimalarathne DN, Bouhari A, Arnous MMR, et al. Evaluation of the diagnostic efficacy of 18 F-fluorine-2-deoxy-D-glucose PET/CT for lung cancer and pulmonary tuberculosis in a tuberculosis-endemic country. Cancer Med. 2020;9:931–42. doi:10.1002/cam4.2770.

13. Hori S, Nishiumi S, Kobayashi K, Shinohara M, Hatakeyama Y, Kotani Y, et al. A metabolomic approach to lung cancer. Lung Cancer. 2011;74:284–92. doi:10.1016/j.lungcan.2011.02.008.

14. Hassanein M, Callison JC, Callaway-Lane C, Aldrich MC, Grogan EL, Massion PP. The state of molecular biomarkers for the early detection of lung cancer. Cancer Prev Res (Phila). 2012;5:992–1006. doi:10.1158/1940-6207.CAPR-11-0441.

15. Ramirez JL, Sarries C, de Castro PL, Roig B, Queralt C, Escuin D, et al. Methylation patterns and K-ras mutations in tumor and paired serum of resected non-small-cell lung cancer patients. Cancer Lett. 2003;193:207–16. doi:10.1016/s0304-3835(02)00740-1.

16. Zhang J, Fu B, Li M, Mi S. Secretome of activated fibroblasts induced by exosomes for the discovery of biomarkers in non-small cell lung cancer. Small. 2021;17:e2004750. doi:10.1002/smll.202004750.

17. Hori S, Nishiumi S, Kobayashi K, Shinohara M, Hatakeyama Y, Kotani Y, et al. A metabolomic approach to lung cancer. Lung Cancer. 2011;74:284–92. doi:10.1016/j.lungcan.2011.02.008.

18. Chen X, Zhou F, Li X, Yang G, Zhang L, Ren S, et al. Folate receptor-positive circulating tumor cell detected by LT-PCR-based method as a diagnostic biomarker for non-small-cell lung cancer. J Thorac Oncol. 2015;10:1163–71. doi:10.1097/JTO.0000000000000606.
19. Jacob M, Lopata AL, Dasouki M, Abdel Rahman AM. Metabolomics toward personalized medicine. Mass Spectrom Rev. 2019;38:221–38. doi:10.1002/mas.21548.

20. Kaddurah-Daouk R, Krishnan KR. Metabolomics: a global biochemical approach to the study of central nervous system diseases. Neuropsychopharmacology. 2009;34:173–86. doi:10.1038/npp.2008.174.

21. Mayerle J, Kalthoff H, Reszka R, Kamlage B, Peter E, Schniewind B, et al. Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis. Gut. 2018;67:128–37. doi:10.1136/gutjnl-2016-312432.

22. Hilvo M, de Santiago I, Gopalacharyulu P, Schmitt WD, Budczies J, Kuhberg M, et al. Accumulated metabolites of hydroxybutyric acid serve as diagnostic and prognostic biomarkers of ovarian high-grade serous carcinomas. Cancer Res. 2016;76:796–804. doi:10.1158/0008-5472.CAN-15-2298.

23. Li Y, Wang L, Ju L, Deng H, Zhang Z, Hou Z, et al. A systematic strategy for screening and application of specific biomarkers in hepatotoxicity using metabolomics combined with ROC curves and SVMs. Toxicol Sci. 2016;150:390–9. doi:10.1093/toxsci/kfw001.

24. Li Z, Yang L, Liu Y, Xu H, Wang S, Liu Y, et al. Anti-inflammatory and antioxidative effects of Dan-Lou tablets in the treatment of coronary heart disease revealed by metabolomics integrated with molecular mechanism studies. J Ethnopharmacol. 2019;240:111911. doi:10.1016/j.jep.2019.111911.

25. Mika A, Wojtowicz W, Ząbek A, Mlynarz P, Chmielewski M, Sledzinski T, et al. Application of nuclear magnetic resonance spectroscopy for the detection of metabolic disorders in patients with moderate kidney insufficiency. J Pharm Biomed Anal. 2018;149:1–8. doi:10.1016/j.jpba.2017.10.037.

26. Expert Committee on tumor markers, laboratory medicine branch of Chinese Medical Association. Basic principles of clinical detection of tumor markers. Chin J of Lab Med. 2004;27:393–8. in Chinese.

27. Hosios AM, Hecht VC, Danai LV, Johnson MO, Rathmell JC, Steinhauser ML, et al. Amino acids rather than glucose account for the majority of cell mass in proliferating mammalian cells. Dev Cell. 2016;36:540–9. doi:10.1016/j.devcel.2016.02.012.

28. Harvie MN, Campbell IT, Howell A, Thatcher N. Acceptability and tolerance of a low tyrosine and phenylalanine diet in patients with advanced cancer – a pilot study. J Hum Nutr Diet. 2002;15:193–202. doi:10.1046/j.1365-277x.2002.00365.x.

29. Kim HJ, Jang SH, Ryu JS, Lee JE, Kim YC, Lee MK, et al. The performance of a novel amino acid multivariate index for detecting lung cancer: A case control study in Korea. Lung Cancer. 2015;90:522–7. doi:10.1016/j.lungcan.2015.10.006.

30. Wei J, Xie G, Zhou Z, Shi P, Qiu Y, Zheng X, et al. Salivary metabolite signatures of oral cancer and leukoplakia. Int J Cancer. 2011;129:2207–17. doi:10.1002/ijc.25881.

31. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25:968–76. doi:10.1038/s41591-019-0458-7.
32. Neurauter G, Grahmann AV, Klieber M, Zeimet A, Ledochowski M, Sperner-Unterweger B, et al. Serum phenylalanine concentrations in patients with ovarian carcinoma correlate with concentrations of immune activation markers and of isoprostane-8. Cancer Lett. 2008;272:141–7. doi:10.1016/j.canlet.2008.07.002.

33. Ploder M, Neurauter G, Spittler A, Schroecksnadel K, Roth E, Fuchs D. Serum phenylalanine in patients post trauma and with sepsis correlate to neopterin concentrations. Amino Acids. 2008;35:303–7. doi:10.1007/s00726-007-0625-x.

34. Huang Y, Xiao M, Ou J, Lv Q, Wei Q, Chen Z, et al. Identification of the urine and serum metabolomics signature of gout. Rheumatology. 2020;59:2960–9. doi:10.1093/rheumatology/keaa018.

35. Stolzenberg-Solomon R, Derkach A, Moore S, Weinstein SJ, Albanes D, Sampson J. Associations between metabolites and pancreatic cancer risk in a large prospective epidemiological study. Gut. 2020;69:2008–15. doi:10.1136/gutjnl-2019-319811.

36. Li J, Li J, Wang H, Qi LW, Zhu Y, Lai M. Tyrosine and Glutamine-Leucine Are Metabolic Markers of Early-Stage Colorectal Cancers. Gastroenterology. 2019;157:257–9. .e5 . doi: 10.1053/j.gastro.2019.03.020.

37. Du Y, Fan P, Zou L, Jiang Y, Gu X, Yu J, et al. Serum Metabolomics Study of Papillary Thyroid Carcinoma Based on HPLC-Q-TOF-MS/MS. Front Cell Dev Biol. 2021;9:593510. doi:10.3389/fcell.2021.593510.

**Figures**

![Non-targeted analysis](image1)

**Training Set**

| group | TB | LC | NC |
|-------|----|----|----|
| serum | 30 | 55 | 35 |

![Targeted analysis](image2)

**Identification Set**

| group | TB | LC | NC |
|-------|----|----|----|
| serum | 63 | 88 | 63 |

**Test Set**

| group | TB | LC | NC | P |
|-------|----|----|----|---|
| serum | 89 | 119| 120| 30 |

**Figure 1**

Sample grouping information.
Figure 2

The quality control sample positive-ion BPI chart.

Figure 3

Volcano map analysis a: LC vs. NC; b: TB vs. NC; c: LC vs. TB The gray point in the figure indicates that the disease group shows no significant difference compared with the non-cancer group; the red point indicates that the disease group is significantly upregulated compared to the healthy group; and the green point indicates that the disease group is significantly downregulated compared with the non-cancer group.
Figure 4

Multivariate statistical analysis of TB, LC, and NC groups (a: TB vs. LC vs. NC PCA analysis; b: TB vs. LC vs. NC PLS-DA analysis; c: TB vs. NC PLS-DA analysis, $R^2_X=0.157$, $R^2_Y=0.916$, and $Q^2=0.692$; d: LC vs. NC PLS-DA analysis, $R^2_X=0.163$, $R^2_Y=0.772$, and $Q^2=0.567$; e, f: Two sets of permutation tests).
Figure 5

Main metabolic pathways a: biosynthesis of phenylalanine, tyrosine, and tryptophan; b: phenylalanine metabolism; c: porphyrin and chlorophyll metabolism; and d: glycerophospholipid metabolism.
Figure 6

The interconnected network between the markers, including lipids, carbohydrates, and amino acids (The green triangle or circle represents the downregulated compared with the non-cancer group, the red triangle or circle represents the upregulated trend compared with the non-cancer group.)
Figure 7

Heat map analysis of biomarkers. (a, b: training set; c, d: identification set)
Figure 8

Results of the ROC curve in the training and identification set. (a, c: L-phenylalanine; b, d: phenylalanylphenylalanine)
Figure 9

Graph of the sample injection results (a: phenylalanylphenylalanine standard; b: phenylalanylphenylalanine peak in blood sample)
Figure 10

The level of phenylalanylphenylalanine in each study group
Figure 11

SVM model modeling accuracy results.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf