BMC Cancer

RESEARCH

Open Access

Metabolic characteristics of benign and malignant pulmonary nodules and establishment of invasive lung adenocarcinoma model by high-resolution mass spectrometry

Junbao Zhang[1,2,4†], Zhihan Zhang[1,2†], Yuying Liu[1], Yanyi Hou[1], Ruifang Pang[1], Yuenan Wang[3*] and Ping Xu[1*]

Abstract

Background Increasing pulmonary nodule presentations in lung adenocarcinoma patients reveal diagnostic limitations of CT-based invasiveness assessment. The critical unmet need lies in developing non-invasive biomarkers differentiating invasive adenocarcinoma from premalignant lesions and benign nodules, while characterizing metabolic trajectory from health to metastatic disease.

Methods Untargeted metabolomics analyzed plasma samples from 102 subjects stratified into four cohorts: confirmed adenocarcinoma ($n=35$), benign nodules ($n=22$), precursor lesions ($n=24$), and healthy controls ($n=21$). Multivariate analysis identified discriminative metabolites for constructing an infiltration prediction model.

Results Three diagnostic groups exhibited distinct metabolic profiles. Hexaethylene glycol, tetraethylene glycol, and Met-Thr showed stage-dependent concentration gradients. Progressive malignancy correlated with elevated levels of 41 metabolites. An eight-metabolite panel achieved AUC 0.933 (0.873–0.994) in distinguishing precursors from early malignancies, sustained through internal validation (AUC 0.934, 0.905–0.966).

Conclusions Met-Thr depletion inversely correlates with malignancy progression, while eight-metabolite signatures demonstrate diagnostic potential for preoperative infiltration assessment in nodular adenocarcinoma.

Keywords Lung nodules, Lung adenocarcinoma, Metabolomics, High-resolution mass spectrometry, Noninvasive metabolomic model

† Junbao Zhang and Zhihan Zhang contributed equally to this work.

*Correspondence:
Yuenan Wang
yuenan.wang@yale.edu
Ping Xu
ping-xu@hotmail.com
Full list of author information is available at the end of the article

Background

As the leading contributor to cancer-related mortality, lung cancer is expected to account for 2.2 million incident cases and 1.8 million deaths globally in 2020. It constitutes the predominant cause of cancer mortality in males and the second leading cause in females, surpassed only by breast cancer [1]. As the world's most populous nation, China confronts distinct challenges in pulmonary oncology management. Demographic aging and nationwide implementation of screening programs have synergistically driven a sustained increase in age-standardized lung cancer incidence rates over the past decade [2, 3]. China will record approximately 871,000 incident lung cancer cases and 767,000 cancer-specific deaths in 2022, constituting 18.1% and 23.9% of total malignant tumor morbidity and mortality, respectively [4]. The elevated mortality rate is primarily attributable to only 30% of cases receiving stage I diagnoses, with over two-thirds first identified at advanced progression. Early-stage manifestations typically involve CT-detectable pulmonary nodules that lack pathognomonic clinical presentations [5]. Studies have shown that the five-year survival rate for stage I lung cancer is 65%, while the survival rate for stage IV lung cancer decreases to 5% [5, 6]. Widespread implementation of standardized screening protocols coupled with early-stage diagnostic interventions constitutes a critical pathway for mortality rate mitigation [7].

With over 50% of all cases being Adenocarcinomas, this is the most common histologic subtype of lung cancer [8]. The American Thoracic Society (ATS), European Respiratory Society (ERS), and International Association for the Study of Lung Cancer (IASLC) collaboratively established a multidisciplinary classification system for lung adenocarcinoma in 2011. This framework categorizes lesions into three principal groups: pre-invasive (encompassing adenocarcinoma in situ [AIS] and atypical adenomatous hyperplasia [AAH]), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC) [9]. The World Health Organization (WHO) established a histopathological classification system for lung adenocarcinoma in 2015, building upon the 2011 IASLC/ATS/ERS framework. This taxonomy was subsequently refined in 2021, with AAH and AIS reclassified as precursor glandular lesions (PGL) [10]. The widespread adoption of low-dose computed tomography (LDCT) screening has significantly increased detection of pulmonary nodules, many demonstrating indolent biological behavior. Clinical management prioritizes balancing surgical intervention timing to prevent overtreatment while ensuring timely therapeutic intervention. Accurate differentiation between benign nodules and early-stage malignant transformation is therefore critical [11].

Differentiating between healthy and malignant lung nodules has become increasingly difficult for doctors in recent years as public awareness of early lung cancer screening has grown [12]. Multinational guidelines currently recommend LDCT as the best noninvasive diagnostic method for tumor detection and evaluation [13, 14]. However, there are issues with overdiagnosis, high false-positive rates, uneven criteria for determining high-risk populations, and cost-effectiveness when using LDCT [15, 16]. Therefore, enhancing lung cancer screening efficacy requires developing early detection biomarkers and optimizing screening protocols. Blood-based biomarkers hold particular promise due to their accessibility, stability, and molecular diversity (encompassing proteins, nucleic acids, exosomes, lipids, and metabolites), which reflect systemic pathophysiological states including tumor-derived signatures [17], It has been extensively utilized in many forms of therapeutic early screening since its composition can disclose the general pathophysiology of tissues and organs [18]. Metabolomics, an emerging discipline following transcriptomics and proteomics, occupies the terminal position in gene regulatory networks, proximate to phenotypic manifestations. This proximity enables direct reflection of biological system functionality. The methodology facilitates disease trajectory prediction through detection of endogenous and environmentally modulated metabolic end-products. Liquid chromatography-tandem mass spectrometry (LC–MS) predominates metabolomic research due to its broad metabolite coverage, high sensitivity, and expansive dynamic range [19, 20]. Thus, this study employs LC–MS-based untargeted metabolomics to identify diagnostic biomarkers differentiating benign and malignant pulmonary nodules in plasma.

Our untargeted metabolomics case–control study systematically characterized metabolic trajectory alterations from health to lung carcinogenesis. Analysis of 102 plasma samples revealed temporal metabolic progression across stratified cohorts: established lung cancer, adenocarcinoma precursor lesions, and benign-nodule controls. An eight-plasma-metabolite panel demonstrated diagnostic accuracy for distinguishing malignant transformation from precursor states. Overall, our results opened the door to more accurate disease detection and therapy by revealing a whole metabolic landscape that spans from healthy populations to lung adenocarcinomas.

Methods

Gathering of clinical samples

This study analyzed 102 plasma specimens collected from Peking University Shenzhen Hospital (August 2021-January 2022), comprising 21 healthy controls(HC),

22 benign pulmonary nodules(BN), 24 adenocarcinoma precursor lesions(PGL), and 35 confirmed lung adenocarcinomas(LC). Demographic and clinical parameters (gender, age, histopathology, medical/medication history) were recorded. Benign nodules encompassed inflammatory lesions, granulomas, lymphoid hyperplasia, bronchiolar metaplasia, and fibrosis. Preoperative fasting participants underwent standardized blood collection: 4–6 mL in K2EDTA tubes, with immediate aliquoting of 1 mL into pre-chilled EP tubes. Following dual centrifugation (1600 g, 10 min; then 1600 g, 30 min, 4°C), supernatants were cryopreserved at -80°C after transfer to barcoded EP tubes.

Untargeted metabolomics investigations

Plasma samples (100μL) underwent biphasic extraction with chilled dichloromethane/methanol (2:1, 400μL) via vortex mixing (10 min, 4°C). Centrifugation (16,000 g, 10 min, 4°C) separated aqueous (upper) and lipid (lower) phases for independent collection. Concentrated extracts were lyophilized and stored at -80°C. For LC–MS analysis, reconstitution involved adding 100μL H2O with vortexing (2 min) and sonication. Lipid phase analysis required 20μL aliquot treatment with isopropanol/acetonitrile/H2O (2:1:1). Chromatographic separation was achieved using HPLC coupled to Q-Exactive HRMS (Thermo) in dual polarity mode, with bioinformatics processing of raw MS data.

Data analysis for metabolomics

Detection of metabolite

High-resolution mass spectrometry was employed for metabolite extraction, with subsequent detection conducted in both positive and negative ion modes. This analytical approach generated substantial mass spectral datasets. Initial data processing involved three sequential phases: (1) format conversion using MS Convert to transform raw data into mzXML format; (2) feature detection through XCMS in R, capturing mass-to-charge ratios (m/z), retention times, and ion intensities; (3) database matching via metaX software against HMDB and KEGG repositories using primary m/z values. The XCMS workflow systematically executed peak alignment, quantification, and chromatographic feature identification. Primary metabolite annotations were established by cross-referencing experimental m/z values with database entries. To resolve structural isomers sharing identical m/z values—a prevalent limitation in primary identification—secondary mass spectral patterns were rigorously compared against authenticated reference spectra. This tandem verification protocol, combining primary m/z matching with MS/MS spectral congruence analysis, significantly enhanced annotation confidence across all detected metabolic features.

Metabolites quantification

Chromatographic peak regions at the substance level yield quantitative metabolite profiles. To ensure data quality, intensity values for each metabolite across all samples were extracted using XCMS. Subsequent preprocessing involved: 1) elimination of low-abundance features (ion detection rates < 50% in quality control samples or < 80% in experimental samples); 2) missing value imputation via K-Nearest Neighbors algorithm; and 3) normalization through Probabilistic Quotient Normalization. Features demonstrating coefficient of variation (CV) exceeding 50% in QC replicates were excluded from subsequent analysis due to excessive intra-experimental variability, as such fluctuations invalidated quantitative variance assessments.

Intergroup comparative analysis

Univariate and multivariate statistical analyses were conducted via the open-source MetaX metabolomics platform to detect group-specific differential metabolites. Methodologies encompassed parametric/non-parametric hypothesis testing, principal component analysis (PCA), fold-change quantification, partial least squares (PLS) regression, and PLS discriminant analysis (PLS-DA) for experimental designs containing three or more biological replicates.

Multi-phenotype metabolic variation analysis

For experimental configurations involving multiple phenotypic groups, comparative metabolic profiling was performed through partial least squares discriminant analysis (PLS-DA) and one-way analysis of variance (ANOVA). Variable Importance in Projection (VIP) scores were computed to quantify each metabolite's contribution to intergroup discrimination, with a VIP threshold ≥ 1.0 applied for biomarker screening.

Multi-comparison group integrative analysis

In studies incorporating multiple phenotypic groups and pairwise comparisons, a hierarchical analytical framework was implemented. This combined global quantitative assessments (e.g., correlation network analysis) with localized qualitative evaluations (e.g., Venn diagram analyses of differential metabolites across comparison sets) to characterize both functional divergences and systemic relationships between experimental groups.

Diagnostic modeling framework

Metabolite biomarker panels were constructed through least absolute shrinkage and selection operator (LASSO) regression using glmnet (v4.1–7, R v4.1.3). Predictive

modeling compared four machine learning architectures: 1) random forest (RF) via randomForest (v4.1–1.2), 2) support vector machine (SVM) using e1071 (v1.7–16), 3) K-nearest neighbors (KNN) implemented with class (v7.3–23), and 4) Gradient Boosting Trees (XGBoost) through xgboost (v1.7.10.1). Data preprocessing and model integration were conducted using caret (v7.0–1), with data manipulation performed via tidyverse (v2.0.0). Diagnostic efficacy was quantified through receiver operating characteristic (ROC) curve analysis using pROC (v1.18.4), with area under the curve (AUC) values calculated to evaluate predictive performance in pulmonary nodule classification.

Results

Metabolomics untargeted metabolic assays

Plasma samples were retrospectively collected from four cohorts: 35 lung cancer cases (including MIA and IAC), 22 benign pulmonary nodule patients, 24 precursor glandular lesion cases (including AAH and AIS), and 21 healthy controls. Demographic characteristics are detailed in Table S1, with Fig. 1 presenting the study design schematic. While no significant intergroup differences in gender distribution were observed ($p > 0.05$), statistically significant age disparities emerged between groups ($p < 0.05$).

Untargeted metabolomic profiling was performed on 102 human specimens (including quality control replicates) using a Q-Exactive high-resolution mass spectrometer (Thermo Scientific) coupled with high-performance liquid chromatography (HPLC) in dual-polarity mode. Mass spectral data acquisition was integrated with bioinformatic interrogation through compound discovery pipelines. Post-extraction metabolites were subjected to high-resolution tandem mass spectrometry (HRMS/MS), yielding final ion statistics presented in Table 1. Initial detection identified 2,516 positive-mode and 6,632 negative-mode metabolic features, with subsequent MS/MS verification confirming 662 positively-charged and 87 negatively-charged molecular species.

Intensity distribution visualizations (box plots and kernel density estimates) were generated to characterize median metabolite abundances and dynamic ranges across samples (Fig. S1). Principal component analysis (PCA), a multivariate dimensionality reduction technique, was applied to derive composite variables capturing maximal metabolic variance. As depicted in Fig. S2, exceptional clustering of quality control samples demonstrated analytical reproducibility, validating methodological reliability.

HMDB (p) denotes the number of metabolite peaks that were matched to primary molecular weights using the HMDB database; KEGG (p) denotes the number of metabolite peaks that were matched to

Table 1 Extraction and identification of mass spectra peak in metabolomics

Mode	All	MS2	HMDB (p)	KEGG (p)	Annotated
negative	2516	662	1380	1454	1653
positive	6632	87	4206	4190	4834

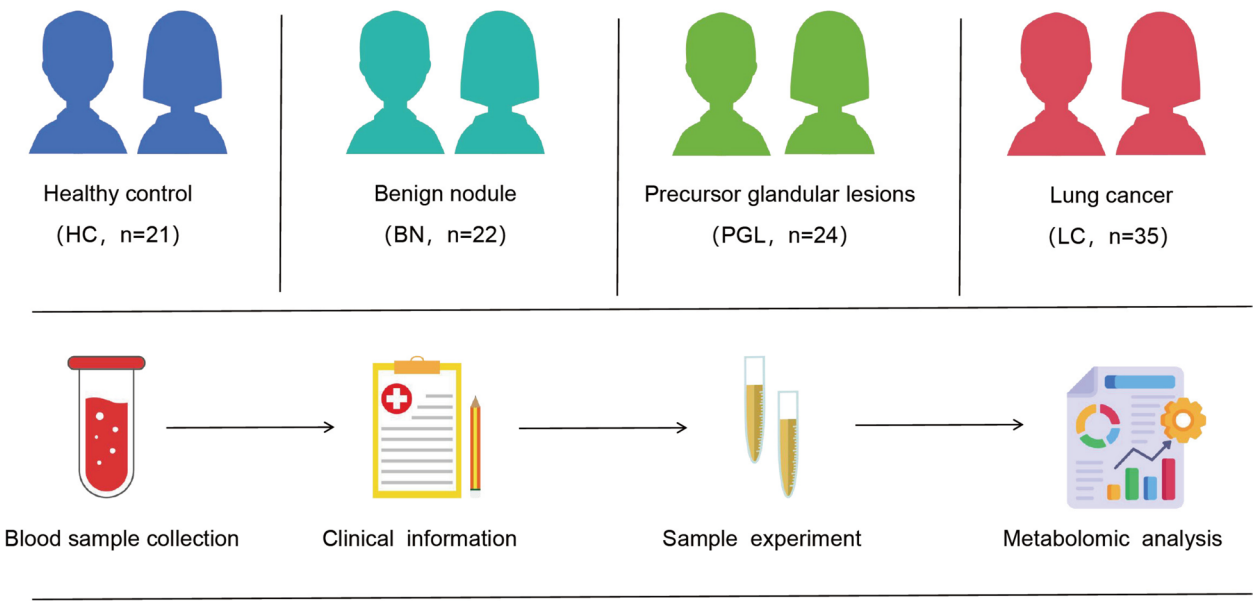

Fig. 1 Overview of the study design

primary molecular weights using the KEGG database; All denotes the number of all metabolite peaks; MS2 denotes the number of metabolite peaks with secondary mass spectrometry identifications; and Annotated denotes the number of peaks with identification information (primary or secondary). metabolite peaks, and Annotated describes the number of peaks that have secondary or primary identifying information.

Metabolite identification

As depicted in Fig. 2A, a comprehensive classification of identified secondary metabolites was performed. Initial categorization organized molecular species into nine major classes: hydrocarbons, homogeneous non-metallic compounds, lignans/neolignans, alkaloid derivatives, phenylcyclic compounds, mixed transition metal complexes, nucleoside analogs, organic

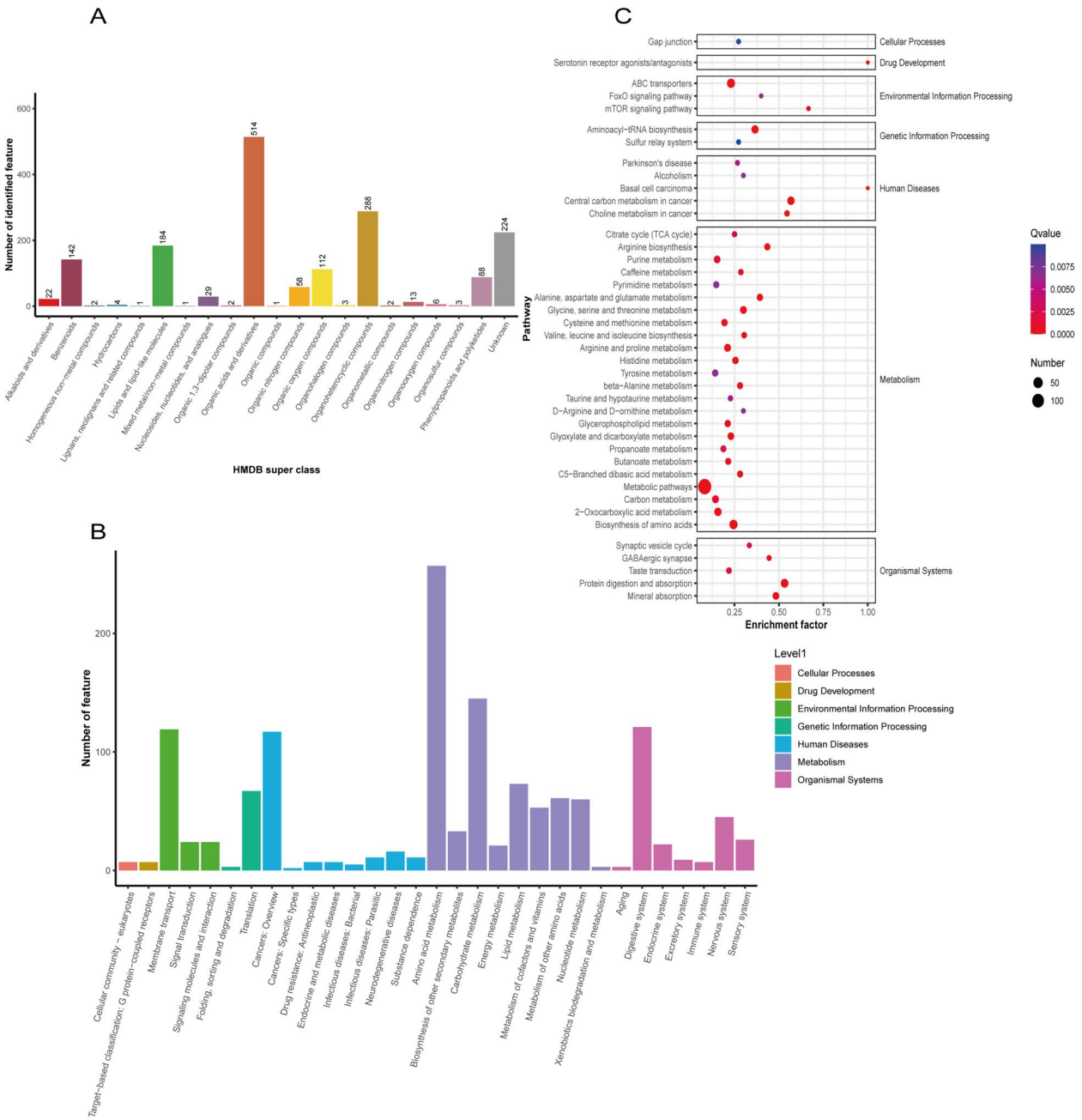

Fig. 2 The quantitative analysis of secondary identification metabolites. **A** Macroscopic evaluation of secondary identification metabolites. **B** Peak number of metabolites participating in KEGG function item (level 2). **C** Participate in enrichment analysis of KEGG function items (the third layer.)

1,3-dipolar compounds, and composite organic molecules (encompassing acid derivatives, nitrogenous/oxygenated/halogenated organics, heterocycles, organometallics, structural compounds, phenylcarbonyl species, and uncharacterized entities). Quantitative analysis revealed organic acid derivatives as the predominant class, representing 514 distinct molecular species.

Metabolic pathway enrichment analysis (Fig. 2B-C) employed KEGG ontology to stratify metabolites into seven functional domains: pharmacological activity, cellular processes, genetic regulation, environmental adaptation, disease pathogenesis, microbial interactions, and metabolic transformation. Systematic quantification demonstrated metabolic processes as the most populous functional category, while disease-associated metabolites – particularly those linked to oncogenesis, cancer subtypes, and infectious pathologies – constituted a minor proportion. Statistical modeling of metabolic networks elucidated compound biosynthetic relationships and their macro-level biological implications.

Statistical and cluster analysis of differential metabolites in multiple phenotype groups

Quantitative characterization of secondary metabolites revealed distinct distribution patterns among the three clinical cohorts (Fig. S3). Differential metabolite identification employed a dual statistical framework: univariate analysis incorporating fold-change ratios and t-test derived p-values (<0.05 threshold), combined with multivariate partial least squares discriminant analysis (PLS-DA Variable Importance in Projection [VIP] >1) as illustrated in Fig. 3A. This integrated methodology identified 788 significantly altered metabolites, including 168 secondary differential metabolites exhibiting differential expression across all three cohorts (Fig. 3B).

Cluster analysis implemented through Mfuzz classified stage-specific metabolites into six expression trajectories (Fig. 3C). The partitioning algorithm divided the sample set ($n=102$) into k mutually exclusive subclasses, where each specimen was assigned to the subclass with minimal Euclidean distance to its centroid. Cluster 1 ($n=41$ metabolites) demonstrated progressive upregulation throughout airway disease progression, dominated by organic acid derivatives (8 species). Cluster 5 ($n=22$ metabolites) exhibited sustained downregulation, predominantly comprising heterocyclic compounds (5 species). PGL expression displayed peak levels in Clusters 4 and 6, while reaching nadir values in Clusters 2 and 3.

Analysis of differential metabolites in the two phenotype groups

Comparative analysis of secondary differential metabolites between pairwise groups identified 59 discriminators between LC and PGL cohorts, 104 between LC and HC + BN cohorts, and 75 between PGL and HC + BN cohorts (Fig. 4A). A Venn diagram revealed four shared metabolites across all three groups (Fig. 4B), from which three unique biomarkers were ultimately retained after removing one redundant compound: Tetraethylene glycol, Met-Thr (Methionine-Threonine), and Hexaethylene glycol. As shown in Fig. 4C, distinct distribution patterns emerged among cohorts. Tetraethylene glycol and Hexaethylene glycol concentrations exhibited marked depletion in the PGL group (fold change: -3.2 and -4.1 respectively) compared to elevated levels in HC + BN controls. Conversely, Met-Thr demonstrated significant accumulation in the PGL cohort (fold change: + 2.8) relative to other groups.

Metabolite screening and modeling of secondary differences between LC and PGL groups

There are clear distinctions in the therapeutic management of the PGL and LC groups, with the former needing case-by-case monitoring and follow-up and the latter needing more extensive treatment. These discrepancies between the two groups raise further concerns. Because it can influence clinicians' decisions, identifying the metabolic differentiators between the two groups is especially crucial. Thus, as illustrated in Fig. 5A and B, we created volcano maps, thermograms, and metabolic route maps for each of the top ten differentials to investigate the distinct metabolite profiles and metabolic pathways between the two. Remarkably, we discovered that Met-Thr was also ranked among them, indicating that it was expressed more highly in PGL than in LC and possibly serving as a protective metabolite. The malignancy of lung cancer was linked to the decline in its level. In addition, by conducting KEGG functional enrichment analysis on the metabolic differentiators, as shown in Figs. S4 and S5, we found that the they were mainly related to metabolism. Three metabolites have also been found to be associated with cancers.

Additionally, we generated AUC curves for 59 secondary difference metabolites in the LC and PGL groups to examine the diagnostic efficacy of various tumor markers in differentiating the two groups to explicitly separate them. We were unable to locate any compounds with an AUC of more than 0.8, and only 28 metabolites had a diagnostic efficacy greater than 0.7, indicating that a single metabolite was not very effective in differentiating

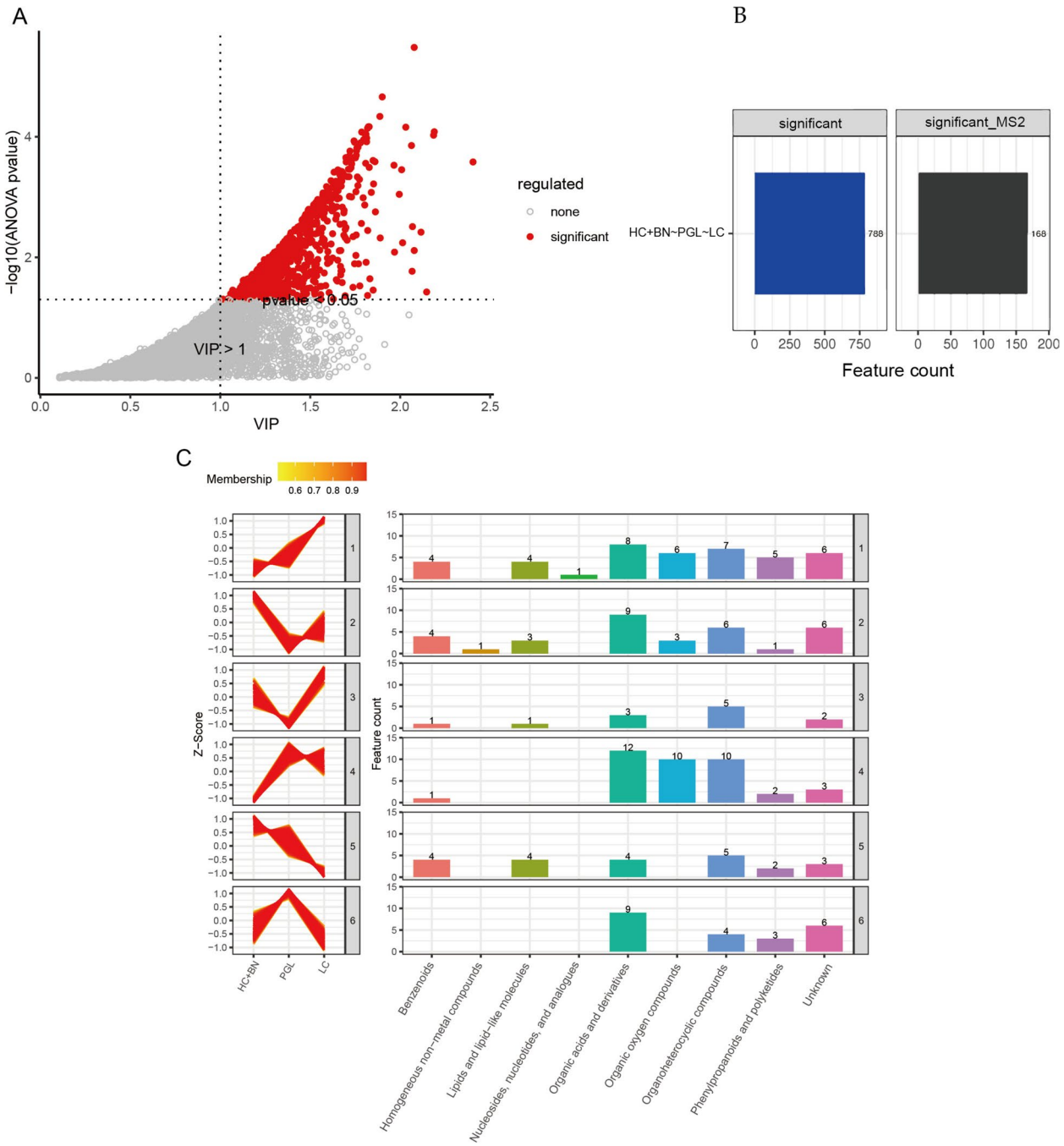

Fig. 3 Statistical and cluster analysis of differential metabolites in multiple phenotype groups. **A** The method of screening differential metabolites(ANOVA *p*-value < 0.05 and PLSDA VIP > 1). **B** The result of screening differential metabolites. **C** Cluster analysis and display of three groups of secondary differential metabolites (K-means)

between the two groups. Because of this, we selected a set of compounds with the goal of establishing a model that would have high credibility and clinical application. The lasso method was employed to screen for differential secondary metabolites in Fig. 5C. A total of eight secondary differential metabolites were screened; the pertinent data is displayed in Table 2. The diagnostic model was then established and its ROC curve was plotted. The diagnostic efficacy of the model was 0.933 (95% CI: 0.873–0.994)) and its internal efficacy was confirmed by resampling the data 1000 times using the bootstrap method. The model's exceptional diagnostic efficacy was fully proved by

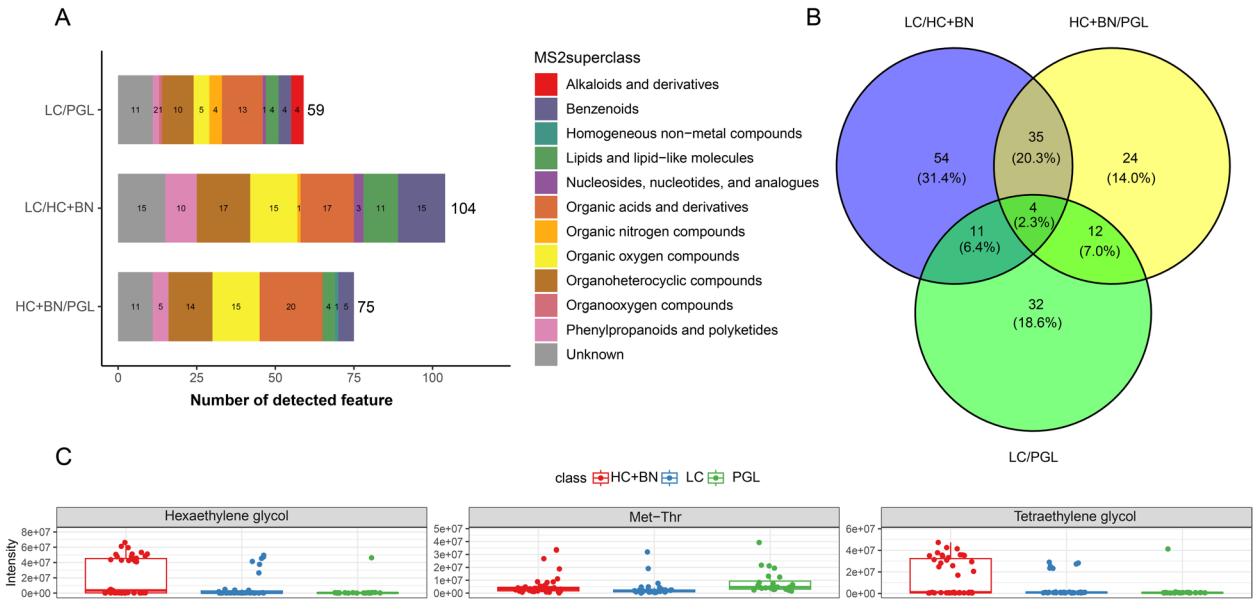

Fig. 4 Analysis of differential metabolites in the two Phenotype groups. **A** Quantitative analysis of secondary differential metabolites in two groups. **B** Wayne diagram based on secondary differential metabolites in each group. **C** Box diagram of the distribution of three different metabolites between three groups

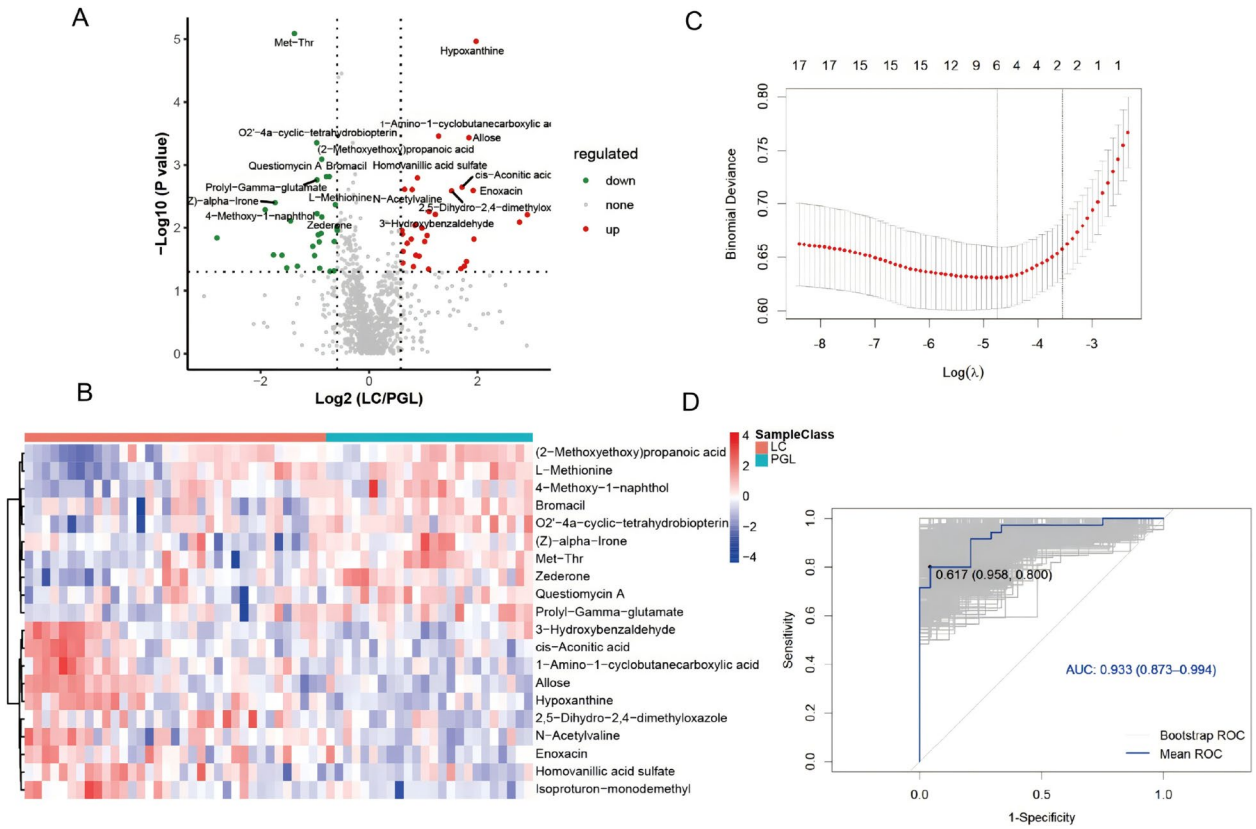

Fig. 5 Secondary differential metabolites in PGL and LC. **A** VolcanoPlot (top 10, log2 greater than 0 is high expression, and log2 less than 0 is low expression metabolite). **B** Heatmap (top 10, Each row is a metabolic peak, and each column is a sample. The color of the heat map is the relative content of each metabolite in each sample). **C** Internal cross-validation of model based on LASSO. **D** Area under the curve of model based on Bootstrap

Table 2 Characteristic variables of the model

Metabolite	ratio	p.value	VIP	cv	regulated	MZ
2,5-Dihydro-2,4-dimethyl oxazole	2.15	0.01	2.94	0.07	up	100.08
Hypoxanthine	3.92	< 0.01	2.45	0.26	up	137.05
Acetyl-2,4-dimethyl oxazole	0.50	0.03	1.72	0.05	down	140.07
N-Acetylvaline	1.73	< 0.01	1.77	0.17	up	160.10
Questiomycin A	0.58	< 0.01	2.04	0.11	down	213.06
O2'-4a-cyclic-tetrahydrobiopterin	0.51	< 0.01	2.20	0.18	down	240.10
Prolyl-Gamma-glutamate	0.51	< 0.01	1.94	0.11	down	244.13
Met-Thr	0.39	< 0.01	3.18	0.00	down	251.11

its average diagnostic efficacy of 0.934 (95% CI: 0.905–0.966) in Fig. 5D, which can assist doctors in differentiating between glandular precursor lesions and lung cancer.

Distinct therapeutic management protocols between PGL and LC cohorts—requiring surveillance monitoring for the former versus comprehensive intervention for the latter—underscore the critical importance of elucidating metabolic disparities. As depicted in Fig. 5A and B, we generated volcano plots, heatmaps, and metabolic pathway maps for the top ten differential metabolites to systematically characterize their divergent metabolic profiles and pathway alterations. Notably, Met-Thr exhibited elevated expression in PGL compared to LC cohorts, suggesting potential protective properties, while its depletion correlated with malignant progression in pulmonary malignancies. KEGG functional enrichment analysis of differential metabolites (Figs. S4-S5) revealed predominant associations with metabolic processes, with three metabolites demonstrating oncological relevance. Subsequent evaluation of diagnostic efficacy through AUC analysis for 59 secondary differential metabolites identified 28 compounds with AUC > 0.7, though none exceeded 0.8, indicating limited discriminatory power of individual biomarkers.

To enhance diagnostic precision, we implemented LASSO regression to screen metabolites (Fig. 5C), ultimately selecting eight biomarkers (Table 2) for model construction. The diagnostic model demonstrated robust performance with an AUC of 0.933 (95% CI: 0.873–0.994), further validated through bootstrap resampling (1000 iterations) yielding a mean AUC of 0.934 (95% CI: 0.905–0.966) (Fig. 5D), effectively differentiating PGL from LC. Comparative machine learning analysis (RF, SVM, KNN, XGBoost) revealed method-specific strengths: KNN achieved optimal accuracy/F1-scores (0.89/0.91), while SVM exhibited near-perfect ROC-AUC (0.98) (Fig. S6A-B). Notably, RF and XGBoost classifiers demonstrated perfect training AUC (1.0), suggesting potential overfitting requiring external validation. Surprisingly, Critical concordance emerged in feature selection: RF incorporated all eight LASSO-identified biomarkers, while XGBoost included seven (Fig. S6C-D). This convergence substantiates the reliability of our LASSO-derived biomarkers. Given overfitting concerns with alternative methods, LASSO regression was retained as the optimal feature selection approach.

The term "ratio" denotes the average ratio between the LC and PGL groups, while "*p*.value" denotes the result of the t-test conducted to compare the two groups. Variable projection importance, or VIP for short, is a measure of how much each metabolite's expression pattern influences and explains how each group of samples is categorized to help with metabolic marker screening (VIP ≥ 1.0 is typically taken as a screening condition). CV stands for coefficient of variation. The mass-to-charge ratio is denoted by MZ.

Discussion

Metabolomics investigates cellular metabolic intermediates and products, influenced by both endogenous and exogenous factors, that regulate physiological and pathological cellular processes. The metabolome provides a phenotypic assessment of systemic and cellular homeostasis, offering translational potential for personalized medicine, pharmacological response evaluation, disease mechanism elucidation, and biomarker discovery [21]. Untargeted metabolomics revealed lung cancer-specific metabolic signatures through comparative plasma analysis of four cohorts: lung adenocarcinoma patients, precursor lesion cases, healthy controls, and benign nodule subjects. Demographic parameters (age/sex) significantly correlated with lipid profiles, amino acid derivatives, and energy metabolism intermediates [22]. Age distribution exhibited significant intergroup variation, potentially reflecting the established oncogenic predisposition associated with advanced age. While this observation

aligns with clinical epidemiology, it necessitates methodological consideration of age-related confounding effects in metabolomic analyses. Our analytical framework incorporated multivariate comparisons (pairwise and multi-group), hierarchical clustering, and LC–MS-based quantitative metabolite profiling. Validation of secondary differential metabolites enabled KEGG pathway mapping to elucidate their functional roles in metabolic reprogramming.

Through multi-step bioinformatic filtering of intergroup statistical variances and intersectional analysis, we identified three differentially abundant metabolites: hexaethylene glycol, tetraethylene glycol, and methionine-threonine (Met-Thr). Met-Thr demonstrated more pronounced differentiation between precursor lesion (PGL) and lung cancer (LC) cohorts, exhibiting progressive depletion correlating with advancing tumor malignancy.As an essential proteinogenic amino acid requiring dietary intake due to endogenous synthesis incapacity, methionine undergoes cellular processing via the methionine cycle. This pathway generates S-adenosylmethionine (SAM), the universal methyl donor that interconnects critical metabolic networks—including glutathione biosynthesis, nucleotide production, folate metabolism, polyamine synthesis, and transsulfuration pathways—through its central role in epigenetic regulation [23]. Hoffman et al.'s seminal 1976 investigation established tumor growth dependency on methionine through demonstrating tumor cell proliferation restriction upon methionine-to-homocysteine substitution in cell culture systems [24]. Emerging evidence demonstrates tumor cells' metabolic dependence on methionine, termed the Hoffman effect. Recent studies reveal methionine's novel role in autophagy inhibition within cancer stem cells, providing mechanistic insight into this phenomenon. Current models propose multiple hypotheses to elucidate tumor-specific methionine auxotrophy [25]. Threonine, an essential amino acid, demonstrates undercharacterized oncological relevance despite its metabolic indispensability. Emerging evidence indicates that hepatocellular carcinoma exhibits elevated serine/threonine-protein kinase 11 (STK11/LKB1) expression, clinically correlating with accelerated tumor progression and unfavorable prognosis [26]. Furthermore, complementary untargeted lipidomic profiling revealed four distinct lipid species demonstrating intergroup variation, though additional discriminators failed to reach statistical significance. This observation suggests either fundamental similarity in lipidomic landscapes between cohorts or reduced statistical power attributable to the study's modest sample size. Using ANOVA ($p < 0.05$) and PLS-DA (VIP > 1) as selection criteria, we identified 788 distinct metabolites, with 168 secondarily screened metabolites exhibiting intergroup differential expression. Cluster analysis revealed six distinct metabolic profiles, where Cluster 1 metabolites demonstrated malignancy-progressive expression patterns within the lung cancer cohort. Consistent with established literature [27], reduced clustering of specific secondarily differentiated metabolites in precancerous lesions (PGL) was observed, implying stage-specific metabolic vulnerabilities with potential translational value for early lesion surveillance.

Comparative metabolomic analysis between PGL and LC cohorts identified 59 statistically significant differential metabolites. Despite implementing ensemble multi-metabolite prediction models aligned with established methodologies for diagnostic classifier development, individual biomarkers failed to demonstrate sufficient discriminatory power between the groups [28]. LASSO regression identified a diagnostic model incorporating eight differential metabolites (Table 2). Notably, hypoxanthine levels in the LC cohort demonstrated significant elevation compared to the PGL group, corroborating prior targeted metabolomics findings regarding purine metabolism dysregulation in lung carcinogenesis [29]. EGFR mutations may drive oncogenesis in lung adenocarcinoma through upregulation of hypoxanthine phosphoribosyltransferase (HGPRT) in the purine salvage pathway, consequently enhancing purine metabolism [30]. Current literature lacks explicit evidence establishing the association between 2,5-dihydro-2,4-dimethyl oxazole and pulmonary adenocarcinoma. Preclinical evidence indicates that isoxazole-fused compounds demonstrate potent EGFR-targeting activity, suggesting their pharmacological potential as antineoplastic agents in lung cancer therapeutics [31]. The 2,5-dihydro-2,4-dimethyl oxazole in this study was significantly elevated in the LC group, suggesting that it may play an anti-cancer role. Questiomycin A has cytotoxic effects on a variety of cancer cells [32–34]. Questiomycin A exhibited significant depletion in lung cancer cohorts, indicating potential pathogenic implications in disease progression.

While this metabolic model demonstrated robust predictive stability (AUC = 0.934) through 1000-resample validation with comparable diagnostic efficacy (AUC = 0.933), mechanistic relationships between other model components and adenocarcinoma pathogenesis lack definitive characterization, with confounding potential from environmental exposures or comorbidities. The model's discriminative capacity provides clinical utility in stratifying nodules approaching microinvasive transition requiring surgical intervention. Study limitations include unresolved etiological contributions from environmental confounders and distinct microbiome-metabolome

profiles observed between smoking and non-smoking cohorts [22]. Notably, the non-smoking status of all healthy controls contrasted with smoking exposure in partial lung cancer cohorts, potentially introducing confounding effects despite unadjusted statistical analyses. Current clinical practice heavily utilizes established lung cancer biomarkers for medical decision-making. While biomarkers comparisons between our metabolic findings and conventional biomarkers would prove methodologically valuable, our prior big-data analyses revealed that these biomarkers predominantly manifest in advanced-stage malignancies, showing negligible deviations in early-stage carcinomas and precancerous lesions [35]. This investigation specifically focuses on early-stage lung adenocarcinoma and its precursor states. Notably, both cohorts exhibited tumor marker levels within normal ranges, with significant inter-individual variability observed in the cancer group. Given their limited discriminative capacity in this clinical context, direct comparative analysis with existing biomarkers was intentionally omitted. Methodological constraints included limited sample size and absence of targeted metabolomic validation, which may affect result generalizability. Furthermore, mechanistic validation through in vitro and in vivo models remains absent for identified metabolite candidates. Our future investigations will focus on two critical directions: validating the robustness of our diagnostic model through multi-center validation cohorts, and implementing targeted metabolomic assays to facilitate clinical translation following rigorous clinical validation. This dual approach aims to develop reliable biomarkers that can ultimately inform clinical decision-support systems. Concurrent mechanistic investigations into differential metabolite pathways and their functional characterization will be essential to elucidate their biological underpinnings in pulmonary carcinogenesis. Nevertheless, our findings establish foundational evidence supporting metabolomic profiling's clinical utility in early lung adenocarcinoma detection.

Conclusions

Our findings demonstrate distinct plasma metabolome alterations in lung cancer compared to healthy states, with specific metabolite concentrations correlating with disease progression. We identified three high-confidence discriminant metabolites showing significant expression gradients across lung adenocarcinoma, precursor lesions, and healthy/benign cohorts. Furthermore, an eight-metabolite panel revealed significant discriminatory capacity between lung cancers and adenocarcinoma precursor lesions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14253-2.

Supplementary Material 1.

Acknowledgements
Peking University Shenzhen Hospital, for their assistance in providing platform and resources for research.

Authors' contributions
ZJ: conceptualization, methodology, formal analysis, data curation, writing – original draft, visualisation. ZZ: conceptualization, methodology, data curation, software, writing – review & editing. LY: methodology, software. HY: conceptualization, methodology. PR: conceptualization, methodology, WY:writing – review & editing, supervision, funding acquisition, project administration. XP: writing – review & editing, supervision, funding acquisition, project administration.

Funding
The study was Supported by the Natural Science Foundation of Guangdong Province (2023A1515012460). Shenzhen Science and technology innovation Commission foundation (JCYJ20210324105411031); General Program for Clinical Research at Peking University Shenzhen Hospital (No. LCYJ2021022).

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The authors confirm that informed consent was obtained from all subjects in the study. All methods were conducted according to the principles of the Declaration of Helsinki and were approved by The Institutional Review Committee and the Medical Ethics Committee of the Peking University Shenzhen Hospital [2021] No. (413).

Consent for publication
Not applicable. Only aggregated data and/or results are shown in this manuscript.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Pulmonary and Critical Care Medicine, Peking University Shenzhen Hospital, Shenzhen 518034, Guangdong Province, People's Republic of China. [2]Peking University Health Science Center, Beijing, China. [3]Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, USA. [4]Department of Pulmonary and Critical Care Medicine, Huashan Hospital, Fudan University, Shanghai, China.

Received: 21 November 2024 Accepted: 2 May 2025
Published online: 08 May 2025

References

1. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol. 2023;20(9):624–39.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA-Cancer J Clin. 2021;71(3):209–49.

3. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chinese Med J-Peking. 2021;134(7):783–91.
4. Xia C, Dong X, Li H, Cao M, Sun D, He S, Yang F, Yan X, Zhang S, Li N, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chinese Med J-Peking. 2022;135(5):584–90.
5. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, Kramer J, Siegel RL. Cancer treatment and survivorship statistics, 2022. CA-Cancer J Clin. 2022;72(5):409–36.
6. Brustugun OT, Gronberg BH, Fjellbirkeland L, Helbekkmo N, Aanerud M, Grimsrud TK, Helland A, Moller B, Nilssen Y, Strand TE, et al. Substantial nation-wide improvement in lung cancer relative survival in Norway from 2000 to 2016. Lung Cancer. 2018;122:138–45.
7. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A, Bolejack V. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol. 2016;11(1):39–51.
8. Pavlova NN, Zhu J, Thompson CB. The hallmarks of cancer metabolism: Still emerging. Cell Metab. 2022;34(3):355–77.
9. Lee HJ, Lee CH, Jeong YJ, Chung DH, Goo JM, Park CM, Austin JH. IASLC/ATS/ERS international multidisciplinary classification of lung adenocarcinoma: novel concepts and radiologic implications. J Thorac Imag. 2012;27(6):340–53.
10. Minami Y. III. The notable topics of the 5th Edition of WHO Classification for the Thoracic Tumours (2021). Gan To Kagaku Ryoho. 2022;49(8):847–52.
11. Yeh YC, Nitadori J, Kadota K, Yoshizawa A, Rekhtman N, Moreira AL, Sima CS, Rusch VW, Adusumilli PS, Travis WD. Using frozen section to identify histological patterns in stage I lung adenocarcinoma of </= 3 cm: accuracy and interobserver agreement. Histopathology. 2015;66(7):922–38.
12. Chen KN. Commentary: Pay attention to low-risk populations for lung cancer, but cautiously interpret ground-glass nodules screened by low-dose computed tomography scan. J Thorac Cardiov Sur. 2020;160(3):833–4.
13. Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. Nat Rev Clin Oncol. 2021;18(3):135–51.
14. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers JJ, Weenink C, Yousaf-Khan U, Horeweg N, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. New Engl J Med. 2020;382(6):503–13.
15. Adams SJ, Stone E, Baldwin DR, Vliegenthart R, Lee P, Fintelmann FJ. Lung cancer screening. Lancet. 2023;401(10374):390–408.
16. Dresler CM, Evans WK. Breathing life into lung cancer screening trials. J Thorac Oncol. 2022;17(11):1244–6.
17. He B, Huang Z, Huang C, Nice EC. Clinical applications of plasma proteomics and peptidomics: Towards precision medicine. Proteom Clin Appl. 2022;16(6):e2100097.
18. Liu C, Xiang X, Han S, Lim HY, Li L, Zhang X, Ma Z, Yang L, Guo S, Soo R, et al. Blood-based liquid biopsy: Insights into early detection and clinical management of lung cancer. Cancer Lett. 2022;524:91–102.
19. Telu KH, Yan X, Wallace WE, Stein SE, Simon-Manso Y. Analysis of human plasma metabolites across different liquid chromatography/mass spectrometry platforms: Cross-platform transferable chemical signatures. Rapid Commun Mass Sp. 2016;30(5):581–93.
20. Heiles S. Advanced tandem mass spectrometry in metabolomics and lipidomics-methods and applications. Anal Bioanal Chem. 2021;413(24):5927–48.
21. Schmidt DR, Patel R, Kirsch DG, Lewis CA, Vander HM, Locasale JW. Metabolomics in cancer research and emerging applications in clinical oncology. CA-Cancer J Clin. 2021;71(4):333–58.
22. Nightingale Health Biobank Collaborative Group. Metabolomic and genomic prediction of common diseases in 700,217 participants in three national biobanks. Nat Commun. 2024;15(1):10092.
23. Martinez Y, Li X, Liu G, Bin P, Yan W, Mas D, Valdivie M, Hu CA, Ren W, Yin Y. The role of methionine on metabolism, oxidative stress, and diseases. Amino Acids. 2017;49(12):2091–8.
24. Hoffman RM, Erbe RW. High in vivo rates of methionine biosynthesis in transformed human and malignant rat cells auxotrophic for methionine. P Natl Acad Sci USA. 1976;73(5):1523–7.
25. Xin L, Li SH, Liu C, Zeng F, Cao JQ, Zhou LQ, Zhou Q, Yuan YW. Methionine represses the autophagy of gastric cancer stem cells via promoting the methylation and phosphorylation of RAB37. Cell Cycle. 2020;19(20):2644–52.
26. Wang Y, Du X, Wei J, Long L, Tan H, Guy C, Dhungana Y, Qian C, Neale G, Fu YX, et al. LKB1 orchestrates dendritic cell metabolic quiescence and anti-tumor immunity. Cell Res. 2019;29(5):391–405.
27. Nie M, Yao K, Zhu X, Chen N, Xiao N, Wang Y, Peng B, Yao L, Li P, Zhang P, et al. Evolutionary metabolic landscape from preneoplasia to invasive lung adenocarcinoma. Nat Commun. 2021;12(1):6479.
28. Li J, Liu K, Ji Z, Wang Y, Yin T, Long T, Shen Y, Cheng L. Serum untargeted metabolomics reveal metabolic alteration of non-small cell lung cancer and refine disease detection. Cancer Sci. 2023;114(2):680–9.
29. Yao Y, Wang X, Guan J, Xie C, Zhang H, Yang J, Luo Y, Chen L, Zhao M, Huo B, et al. Metabolomic differentiation of benign vs malignant pulmonary nodules with high specificity via high-resolution mass spectrometry analysis of patient sera. Nat Commun. 2023;14(1):2339.
30. Geng P, Ye F, Dou P, Hu C, He J, Zhao J, Li Q, Bao M, Li X, Liu X, et al. HIF-1alpha-HPRT1 axis promotes tumorigenesis and gefitinib resistance by enhancing purine metabolism in EGFR-mutant lung adenocarcinoma. J Exp Clin Canc Res. 2024;43(1):269.
31. Ardhapure SS, Sirsat SB. One-pot synthesis of fused isoxazolo[4′,5′:4,5] thiopyrano[2,3-d]pyrimidines as potent EGFR targeting anti-lung cancer agents. Tetrahedron Lett. 2024;151:155325.
32. Gallo A, Ghilardelli F, Atzori AS, Zara S, Novak B, Faas J, Fancello F. Co-occurrence of regulated and emerging mycotoxins in corn silage: relationships with fermentation quality and bacterial communities. Toxins. 2021;13(3):232.
33. Janic Hajnal E, Kos J, Malachova A, Steiner D, Stranska M, Krska R, Sulyok M. Mycotoxins in maize harvested in Serbia in the period 2012–2015. Part 2: Non-regulated mycotoxins and other fungal metabolites. Food Chem. 2020;317:126409.
34. Machihara K, Tanaka H, Hayashi Y, Murakami I, Namba T. Questiomycin A stimulates sorafenib-induced cell death via suppression of glucose-regulated protein 78. Biochem Bioph Res Co. 2017;492(1):33–40.
35. He J, Wang B, Tao J, Liu Q, Peng M, Xiong S, Li J, Cheng B, Li C, Jiang S, et al. Accurate classification of pulmonary nodules by a combined model of clinical, imaging, and cell-free DNA methylation biomarkers: a model development and external validation study. Lancet Digit Health. 2023;5(10):e647–56.

Publisher's Note