Check for updates

Development of a urine-based metabolomics approach for multi-cancer screening and tumor origin prediction

Xinping Xu[1†], Chunyan Zeng[2†], Bei Qing[3†], Yun He[4†], Guodong Song[5], Jiaojiao Wang[4], Shuqi Yu[4], Tao Zhang[4], Qingyan Wei[4], Li Liu[4], He Wen[4], Junyuan Hu[4], Wei Zhang[2], Yan Li[4*], Youxiang Chen[2*] and Zhenkun Xia[3*]

[1]Jiangxi Institute of Respiratory Disease, The First Affiliated Hospital of Nanchang University, Nanchang, China, [2]The First Affiliated Hospital of Nanchang University, Nanchang, China, [3]The Second Xiangya Hospital of Central South University, Changsha, China, [4]Metanotitia Inc., Shenzhen, China, [5]The Second Hospital of Tianjin Medical University, Tianjin, China

Background: Cancer remains a leading cause of mortality worldwide. A non-invasive screening solution was required for early diagnosis of cancer. Multi-cancer early detection (MCED) tests have been considered to address the challenge by simultaneously identifying multiple types of cancer within a single test using minimally invasive blood samples. However, a multi-cancer screening strategy utilizing urine-based metabolomics has not yet been developed.

Methods: We enrolled 911 cancer patients with 548 lung cancer (LC), 177 with gastric cancer (GC), and 186 with colorectal cancer (CRC), alongside 563 individuals with non-cancerous benign diseases and 229 healthy controls (HC) and investigated the metabolic profiles of urine samples. Participants were randomly allocated to discovery and validation cohorts. The discovery cohort was used for identifying multi-cancer and tissue-specific signatures to build the cancer screening and tumor origin prediction models, while the validation cohort was employed for assessing the performance of these models.

Results: We identified and annotated a total of 360 metabolites from the urine samples. Using the LASSO regression algorithm, 18 metabolites were characterized as urinary metabolic biomarkers and exhibited excellent discriminative performance between cancer patients and HC with AUC of 0.96 in the validation cohort. In comparison with the performance of traditional tumor markers CEA, the screening model performed higher sensitivity across the cancer stages, with a particularly increase in sensitivity among early-stage cancer patients. Moreover, the screening model also exhibited in high classification of cancers from non-cancerous group, comprising with HC and benign disease participants. Furthermore, two non-overlapping metabolic panels were selected to differentiate LC from Non-LC and GC from CRC with the AUC values of 0.87 and 0.83 in validation cohorts, respectively. Additionally, the model accurately predicted the origin of three lethal cancers: lung, gastric, and colorectal, with an overall accuracy of 0.75. The AUC values for LC, GC, and CRC were 0.88, 0.88, and 0.80, respectively.

Discussion: Our study demonstrates the potential of urine-based metabolomics for multi-cancer early detection. The approach offers non-invasive cancer screening, promising widespread implementation in population-based programs for early detection and improved outcomes. Further validation and expansion are needed for broader clinical applicability.

Introduction

Cancer is a leading cause of death worldwide, and it is predicted that global cancer incidence will double by 2070 compared to 2020 (1). It was estimated that in 2022, lung cancer (LC), colorectal cancer (CRC), and gastric cancer (GC) were among the five most common cancers and leading causes of cancer-related deaths in China, representing 22.0%, 10.7%, and 7.4% of new cancer cases, respectively, and contributing to 28.4%, 9.3%, and 10.1% of total cancer deaths (2). Cancer screening aids in the detection of cancer at early stage when treatment is both more effective and cost-efficient. To date, the clinical blood tumor biomarkers are limited, including alpha-fetoprotein (AFP), cancer antigen 19-9 (CA19-9), and carcinoembryonic antigen (CEA), all of which demonstrate low sensitivity. Besides, most of the current screening tests are for single cancer, which may lead to elevated false positive rates when used consecutively (3). Therefore, there is an urgent need for non-invasive screening solution for pan-cancer early diagnosis.

Multi-cancer early detection (MCED) tests have the potential to address these challenges by identifying multiple types of cancer simultaneously within a single test, with numerous novel technologies being investigated (4). For example, a targeted cfDNA methylation-based assay called Galleri test was applied to diagnose 50 different cancers (5). Moreover, the CancerSEEK test integrates genetic mutation signals and protein markers to detect eight types of cancer (6). However, these methods require complex sample processing and high depth coverage of sequencing, which are time-consuming and costly, hindering the population wide screening (7). Previous studies have shown the potential application of metabolomics in diagnosing certain cancers, as cancer is acknowledged as a metabolic disorder (8). Additionally, metabolomics technologies exhibit high reliability and reproducibility, making them promising tools for identifying potential biomarkers in tissues, blood, and urine samples (9). Compared to other biological fluids (tissue and blood), urine has the advantages of being inexpensive, easy to handle, and available in large amounts, without requiring invasive treatments for collection, which could potentially broaden its application (10). Besides, it contains a variety of inorganic salts, organic compounds, and diverse exfoliated cell types, providing rich metabolic information

reflecting the systemic metabolic status (11, 12). Moreover, urine as a carrier for blood wastes, contains various metabolic products from pathways (13). Hence, the application of urine-based metabolomics in cancer screening is poised to become a powerful tool for clinical cancer diagnosis. Currently, urinary metabolomics has been applied to identify potential markers for the diagnosis of various cancers. For instance, meta-analysis characterized eleven relevant urinary metabolites for colorectal cancer diagnosis (14). In addition, Mathé et al. (15) and Chan et al. (16) revealed specific urinary metabolic biomarkers of diagnosing lung cancer and gastric cancer, respectively. This demonstrated the clinical potential of urinary metabolomics for cancer diagnosis. To the best of our knowledge, however, a comprehensive multi-cancer screening strategy utilizing urine-based metabolomics has not been developed yet.

Herein, the aim of this study is to develop a urinary metabolomics-based approach for multi-cancer diagnosis by identifying cancer-specific urinary biomarkers, enabling the construction of screening models that can distinguish between cancerous and non-cancerous individuals and accurately predict the origin of three lethal cancers (including LC, CRC, and GC). The urinary metabolomics holds promise of becoming a universally applicable, straightforward, and cost-effective means of early detection for these three cancers across large populations.

Methods

Study population

A case-control study was conducted between September 2018 and January 2020 at three independent centers, including the First Affiliated Hospital of Nanchang University, the Second Xiangya Hospital of Central South University, and the Second Hospital of Tianjin Medical University. Participants included individuals with LC, GC, CRC, non-cancer benign diseases (NCD) of the lung, stomach, and colorectum, as well as healthy controls (HC). The study was approved by the local Institutional Review Board, adhering to the guidelines of the International Conference on Harmonization for Good Clinical Practice and the Declaration of Helsinki, with formal consent obtained from the participants. All

participants were screened as described previously (17). Morning fasting urine samples were collected, centrifuged, and processed using established methods (17).

Sample preparation and metabolites detection

The extraction buffer, consisting of methanol, methyl tert-butyl ether, water, and a mixture of chemical standards (gibberellic acid A3: 0.45 µg/mL, ^{13}C sorbitol: 1 µg/mL, and PE (17:0/17:0): 1 µg/mL), was combined with the samples and sonicated for 15 minutes at 4°C. Following this, 350 µL of methanol/water (v/v, 1:3) was added to facilitate phase separation. The resulting upper lipophilic phase and hydrophilic phase were collected separately after high-speed centrifugation (18000 g, 5 minutes at 4°C, Centrifuge 5430R, Eppendorf, Germany). All aliquots were dried and stored at -80°C for further analysis. The polar fraction was derivatized according to Lisec et al. (18). Metabolite analysis was performed using gas chromatography and UPLC-MS (Waters ACQUITY ultra performance liquid chromatography (UPLC) system coupled to Thermo-Fisher Q-Exactive mass spectrometers with an electrospray ionization (ESI) source). The resuspended samples were injected into a Waters ACQUITY FTN autosampler set to a temperature of 10°C. Data were acquired in both positive and negative modes, with parameters as described previously (17).

Data processing

The raw data (.raw) were initially processed using Metanotitia Inc.'s in-house developed software PAppLine^{TM}. This process involved peak picking, baseline correction, alignment, removal of isotopic peaks, annotation, and baseline noise removal. Metabolic features detected in less than 80% of the subject samples were discarded to mitigate the impact of noise and outliers in the data. Missing values were addressed separately for the subject samples and the quality control samples (QC$_{Nist}$), using the same strategy batch-by-batch. The imputation method varied depending on the missing rate and average intensity of each feature within the batch. For features with a low missing rate and high average intensity, the MICE forest algorithm was applied. For those with a high missing rate or low average intensity, the half-minimum method was used for imputation. Meanwhile, features with high missing rate or low averaging intensity were imputed by the half minimum strategy. The MSTUS (total useful MS signals) method is used to normalize mass spectrometry data to eliminate dilution effects present in urine samples (19). Subsequently, normalization was conducted utilizing the QC-based deep learning method, NormAE algorithm. Finally, a calibration procedure, including logarithm transformation and scaling, was employed.

Feature selection and model construction

Feature selection was independently conducted for each platform within the discovery cohort using the least absolute shrinkage and selection operator (LASSO) algorithm. The LASSO method regularizes parameters of a linear regression model by reducing some coefficients to zero, allowing the selection of features with nonzero coefficients (20). To enhance its stability, we applied an ensemble learning approach, constructing and amalgamating multiple selection systems with a recurring random data-splitting strategy. Subsequently, selected features from the three distinct MS-based platforms were integrated for addressing the forthcoming classification problem. For the training model, a five-fold cross-validation procedure repeated five times was applied to each classifier. The discovery set was divided into two subsets at a ratio of 4:1 to compute the training error on the resulting 20% dataset. Through five repetitions of this process, the average training error was calculated and employed for training hyper-parameters. The model comprises the Balanced Support Vector Machine (SVM) algorithm, a fundamental classification algorithm. These techniques are adept at handling high-dimensional mass spectrometry-based data, particularly in situations with limited sample sizes. To tackle multi-class classification, we used the One vs. Rest approach for these three distinct cancer types and employed ensemble learning techniques like bagging and stacking to enhance the metabolic panel (MP) model. For the test model, parameters were determined during training, and the validation set was used to compute validation error, including confusion matrix, ROC curves, and AUC, for all corresponding classifiers.

Statistical analysis

The data's statistical significance was assessed at a 95% confidence level ($p < 0.05$) using binomial distribution. Data processing and machine learning were conducted in Python (version 3.10.12), leveraging the Numpy (version 1.23.5) and Pandas (version 1.5.3) libraries. Imputation was carried out with Miceforest (version 5.6.2), while feature selection and machine learning employed Scikit-learn (version 1.2.2) and Scipy (version 1.10.1). Matplotlib (version 3.8.2) was employed for generating the confusion matrix to summarize classification results and plot ROC curves. The confusion matrix for multiple classification (among three cancers: GC, CRC, and LC) was generated using Scikit-learn metrics. The heatmap was drawn using R (version 4.2.0). To assess the accuracy, sensitivity, and specificity of the different groups, ROC curves were applied for various scenarios (discovery set, validation set). ROC curves and AUC values were performed by Scikit-learn.

Results

Characteristics of participants

This study recruited participants with three common types of cancer: LC (n = 548), GC (n = 177), and CRC (n = 186), along with 563 patients with non-cancerous benign diseases (NCD) and 229 healthy controls (HC). All cancer patients were diagnosed by histopathology and/or imaging techniques, depending on cancer types, and had not received any prior treatment at the time of urine collection. The cancer, non-cancer disease, and HC groups were randomly assigned to the

discovery and validation cohorts (Table 1). The discovery cohort was utilized for profiling multi-cancer- and tissue-specific signatures and constructing machine learning algorithms, whereas the validation cohort was exclusively employed for evaluating the performance of the machine learning models (Figure 1).

The discovery cohort included 713 cancer patients (431 LC, 142 GC, and 140 CRC), 422 NCD patients, and 172 healthy controls. The gender ratios were comparable between the cancer, NCD, and HC groups ($p > 0.05$), while the cancer patients were found to be the oldest than the other two groups with a median age of 62 ($p < 0.05$) (Table 1). To better establish an early cancer screening model, we enrolled numerous early-stage cancer patients, with 30% of them diagnosed at stage I. The validation cohort comprised of 198 cancer patients, including 117 LC, 35 GC, and 46 CRC, 141 NCD patients, and 57 HC. Consistent with the discovery cohort, the gender distribution was comparable between the cancer, NCD, and HC groups. Furthermore, the cancer group had a higher median age (63 years) compared to the other two groups (53 and 50 years) ($p < 0.05$). The distribution of cancer patients across each stage closely resembled that of the discovery cohort (Table 1).

Establishment of metabolites-based signature panel for pan-cancer detection

After peak alignment and removal of missing values, 125, 13, and 222 metabolites were identified and annotated on the polar, lipid, and GC platforms, respectively, including fatty acids, carbohydrates, amino acid and derivatives, nucleotides and derivatives, acetyl-CoA, diacylglycerol and others. The metabolic reprogramming of cancer cells has extensively been examined in prior studies, shedding light on the alterations in patients' metabolism (8). The 18 metabolites for the discrimination of pan-cancer and non-cancer group have been selected using the LASSO regression algorithm. KEGG pathway enrichment analysis on the metabolites panel uncovered a variety of disrupted metabolic pathways including the starch and sucrose, galactose, and tryptophan metabolism ($p < 0.05$, Figure 2A), which has been well characterized in cancer patients in previous research (21, 22). Subsequently, an SVM algorithm was employed to establish the screening model based on metabolic panel selected from the discovery dataset (named as MP-SVM), and the performance was further evaluated in the validation dataset. Impressively, the MP-SVM model achieved an area under the receiver operating characteristic (AUC) of 0.98 (95% confidence interval (CI): 0.97-1.00) and 0.96 (95% CI: 0.93-0.98) for distinguishing cancer group from HC in the discovery and validation datasets, respectively (Figure 2B). To evaluate the screening advancement of the MP-SVM model, we compared its sensitivity with that of the established clinical tumor biomarker CEA in the validation cohort. The MP-SVM model outperformed the traditional tumor marker CEA. The detection rate of MP-SVM model in the pan-cancer group was higher than that of CEA with specificity > 99% in stages I, II, III, and IV (Figure 2C).

Moreover, the panel exhibited an AUC of 0.83 for discerning cancer group from the non-cancerous group including the HC and NCD patients in the validation cohort (Figure 3A). The detection rates

TABLE 1 Baseline characteristics of the participants.

Variable	Discovery Cohort (N=1307)				Validation Cohort (N=396)			
	Cancer (N = 713)	NCD (N = 422)	HC (N = 172)	p value	Cancer (N = 198)	NCD (N = 141)	HC (N = 57)	p value
Age, median (range)	62 (21-93)	56 (18-87)	47 (18-67)	0.0006	63 (27-82)	53 (19-88)	50 (23-66)	0.0016
Gender, N (%) Female	248 (34.8%)	173 (41%)	92 (53.5%)	0.4793	70 (35.4%)	61 (43.3%)	30 (52.6%)	0.4784
Male	465 (65.2%)	249 (59%)	80 (46.5%)		128 (64.6%)	80 (56.7%)	27 (47.4%)	
Stage, N (%) I	214 (30%)				60 (30.3%)			
II	80 (11%)				19 (9.6%)			
III	123 (17%)				36 (18.2%)			
IV	107 (15%)				31 (15.7%)			
Non-metastasis with unknown staging information	189 (26.5%)				52 (26.3%)			

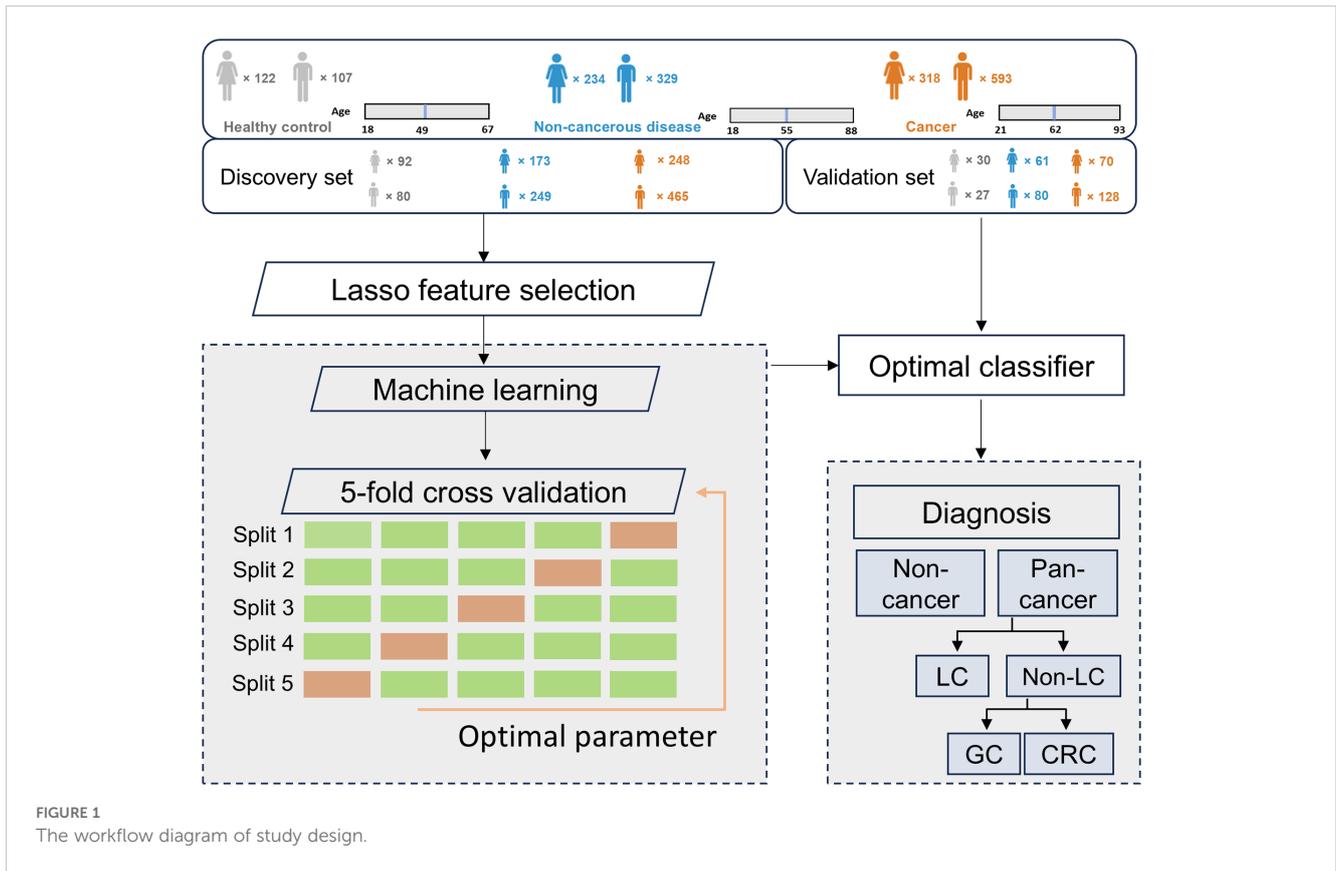NCD, non-cancerous disease; HC, healthy control.

FIGURE 1
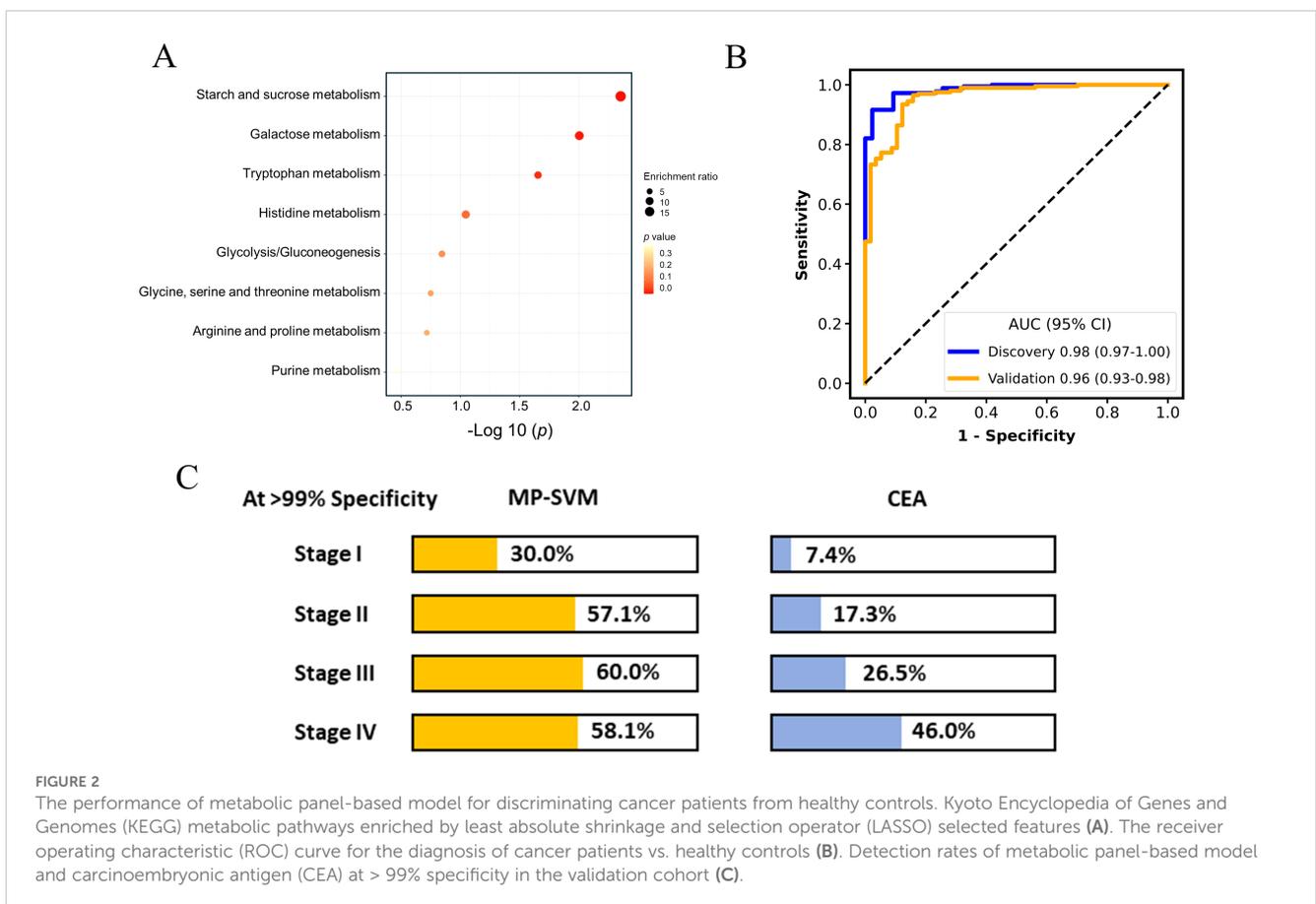The workflow diagram of study design.

FIGURE 2
The performance of metabolic panel-based model for discriminating cancer patients from healthy controls. Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways enriched by least absolute shrinkage and selection operator (LASSO) selected features **(A)**. The receiver operating characteristic (ROC) curve for the diagnosis of cancer patients vs. healthy controls **(B)**. Detection rates of metabolic panel-based model and carcinoembryonic antigen (CEA) at > 99% specificity in the validation cohort **(C)**.
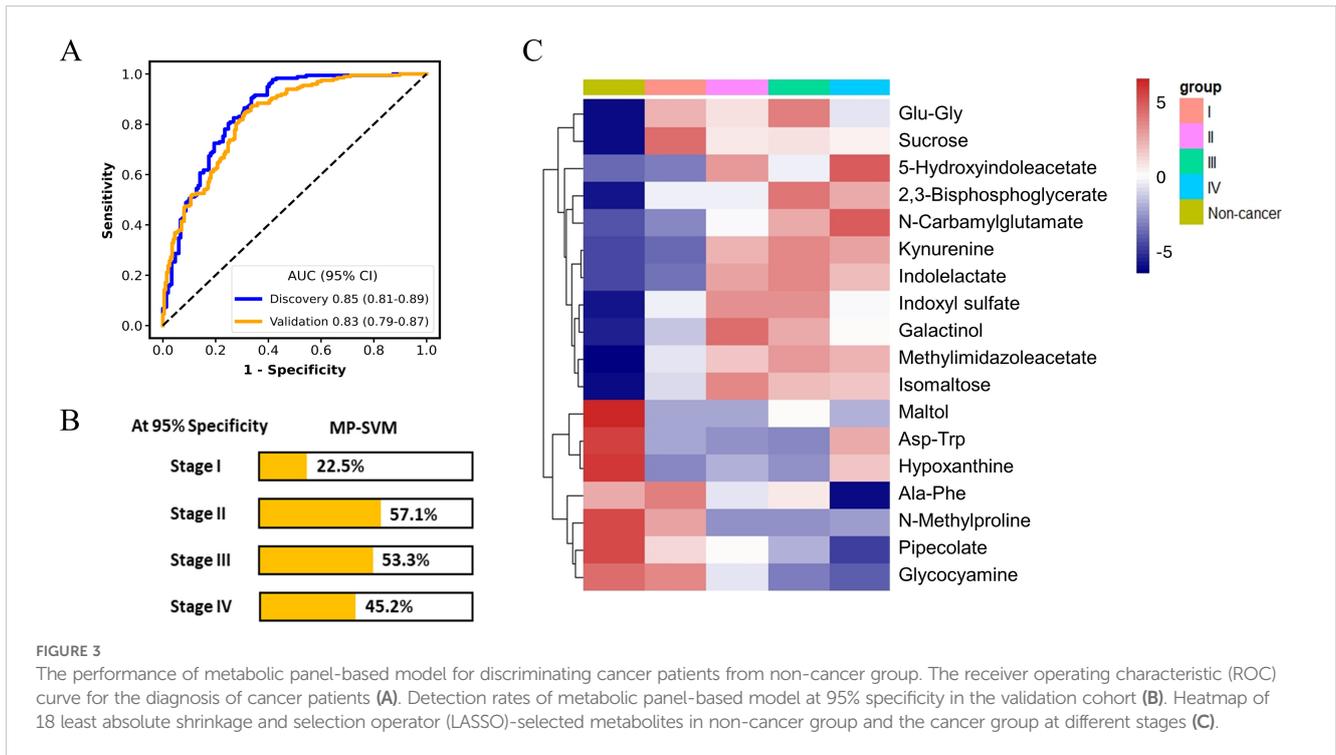
FIGURE 3
The performance of metabolic panel-based model for discriminating cancer patients from non-cancer group. The receiver operating characteristic (ROC) curve for the diagnosis of cancer patients (A). Detection rates of metabolic panel-based model at 95% specificity in the validation cohort (B). Heatmap of 18 least absolute shrinkage and selection operator (LASSO)-selected metabolites in non-cancer group and the cancer group at different stages (C).

of the MP-SVM model in the pan-cancer group with a specificity of 95% for stages I, II, III, and IV were 22.5%, 57.1%, 53.3%, and 45.2%, respectively (Figure 3B). The heatmap was applied to delineate the metabolite trajectories (LASSO-selected metabolites) in stage I, II, III, and IV (Figure 3C). The distinct variations of these metabolites were observed across the four stages of cancer, reflecting intricate shifts in metabolic dynamics throughout cancer progression. More specifically, the changes in most metabolite levels in stage I were not pronounced (e.g., 5-Hydroxyindoleacetate, N-Carbamylglutamate, kynurenine, and indolelactate). And the levels of certain metabolites in stages II and III significantly increased (e.g., kynurenine, indolelactate, indoxyl sulfate, galactinol, methylimidazoleacetate, and isomaltose). The 5-Hydroxyindoleacetate and N-Carbamylglutamate in stage IV showed the most significant increase, while pipecolate and glycocyamine exhibited the most notable decrease. Next, we assessed the potential confounding effect of age on our model by examining the correlation between the model prediction scores and the participants' ages. The results exhibited no significant correlation (R = 0.05, P = 0.32), indicating that age differences are unlikely to impact the accuracy of our model (Supplementary Figure S1).

Establishment of cancer classification model for tumor origin prediction

The accurate prediction of the tumor origin is essential in early cancer detection, as it guides subsequent diagnostic procedures and treatment decisions (23). Given that both GC and CRC originate in the digestive tract, they have been reported to exhibit certain shared metabolic characteristics (24). Thus, in this study, the feature selection process for the cancer classification model consisted of

two stages: distinguishing between LC and non-LC samples (GC and CRC), and then differentiating between GC and CRC samples. Seventeen metabolites were selected by LASSO algorithm to differentiate LC from Non-LC, achieving AUC values of 0.88 (95% CI: 0.83-0.93) and 0.87 (95% CI: 0.82-0.91) in the discovery and validation datasets, respectively (Figure 4A). Among these metabolites, metabolites such as 3-Hydroxybutanoate, cadaverine, N-Cinnamylglycine, and anabasine exhibited upregulation in the non-LC group, whereas 3-Ureidopropionate, O-Acetylhomoserine, 4-Hydroxyphenylpyruvate, xanthurenic acid, N-Carbamylglutamate, 3-Methyl-2-pentenedioic acid, methionyl-aspartate (Met-Asp), and 3-Methylhistidine were found to elevate in the LC group (Figure 4B). These 17 metabolites were further mapped to the KEGG metabolic pathways for pathway analysis. Histidine metabolism, glutathione metabolism, tyrosine metabolism, and phenylalanine, tyrosine, and tryptophan biosynthesis were the most significantly changed pathway when comparing LC and non-LC groups (p < 0.05) (Supplementary Figure S2A).

An additional set of 21 metabolites were selected for classifying GC and CRC, without overlapping with features selected for the LC and non-LC classification. The AUC for the MP-SVM model in the discovery and validation datasets were 0.91 (95% CI: 0.84-0.97) and 0.83 (95% CI: 0.74-0.91), respectively (Figure 4C), with acetylhistamine, trisaminol, and sucrose being the three most significant contributing metabolites (Figure 4D). The 21 metabolites were enriched in 17 metabolic pathways, with 5 pathways (galactose metabolism, carbohydrate metabolism, fructose and mannose metabolism, neomycin, kanamycin, and gentamicin biosynthesis, arginine, and proline metabolism) exhibiting significant differences (p < 0.05) (Supplementary Figure S2B). The two metabolites' panels were pooled together, forming the basis for constructing the tumor origin
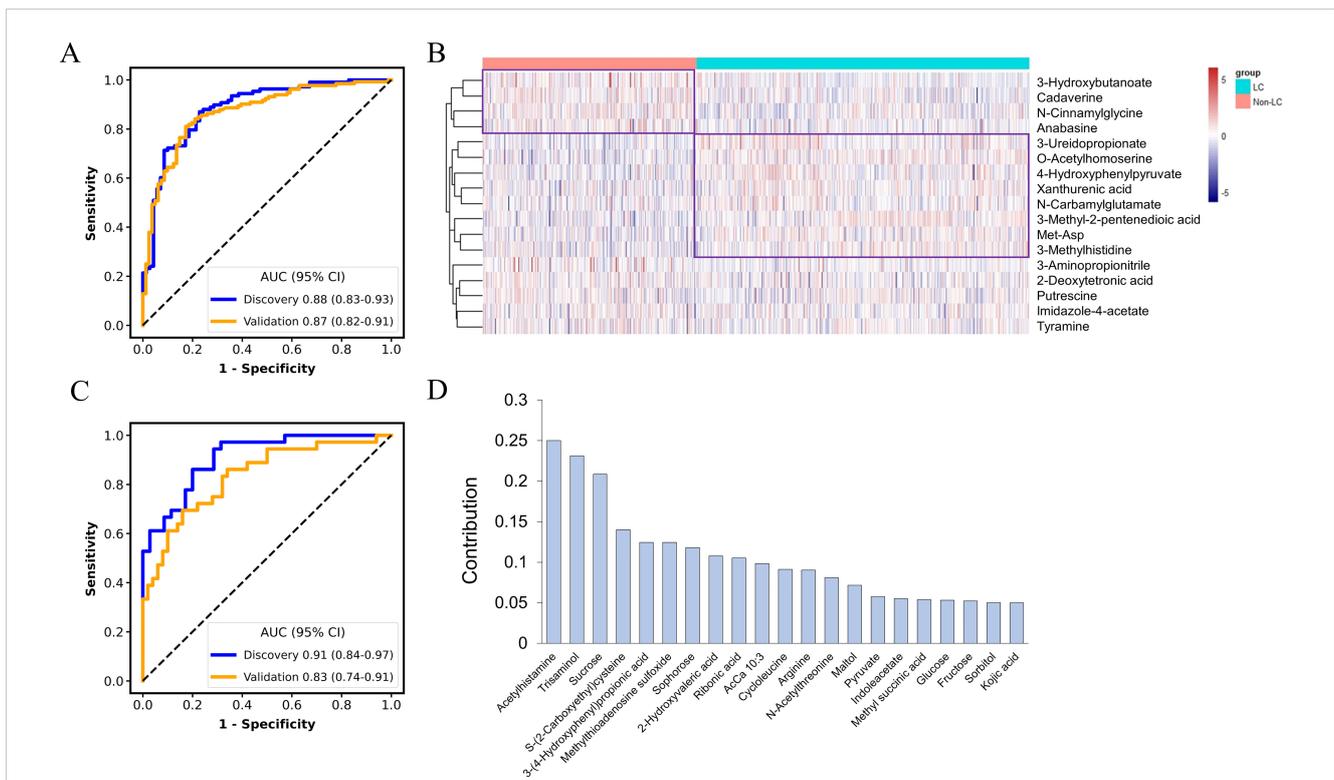
FIGURE 4

Classification of cancers by metabolic panel-based model. The receiver operating characteristic (ROC) curve for the discrimination of lung cancer (LC) patients from non-lung cancer patients **(A)**. Heatmap of the least absolute shrinkage and selection operator (LASSO)-selected metabolites between LC and non-LC patients **(B)**. The ROC curve for the discrimination of gastric cancer patients (GC) from colorectal cancer patients (CRC) **(C)**. Contribution of the selected metabolites to the discrimination model for GC vs. CRC **(D)**.

classification model. In the validation cohort, the model yielded an accuracy of 0.75, with AUC of 0.88 (95% CI: 0.83-0.93) for LC, 0.88 (95% CI: 0.81-0.93) for GC, and 0.80 (95% CI: 0.72-0.87) for CRC, respectively (Figure 5A). Among the three cancer types, LC showed the highest detection rate (0.83), while CRC exhibited the low detection rate (0.52) (Figure 5B).

Discussion

In this study, we developed a novel urine metabolites-based model and revealed urinary metabolic biomarkers exhibiting high discriminative capacity for multi-cancer early detection. Despite the wide range of genetic changes observed in various types of cancer,
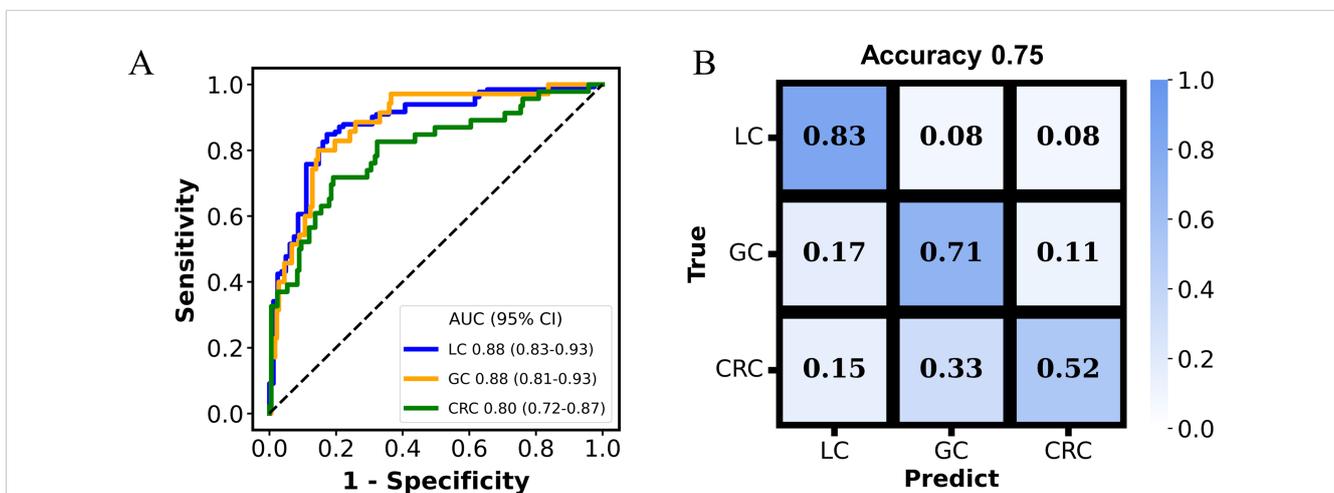

FIGURE 5

The performance of the multi-cancer classification model. The receiver operating characteristic (ROC) curves evaluating the model in discriminating tumor origins in the validation cohort **(A)**. Confusion matrix summarizing the cancer classification results in the validation cohort **(B)**.

there is a possibility that cancer development could impose similar requirements on cellular metabolism, regardless of its tissue of origin (25). Besides, subtle distinctions exist among cancers originating from different sites, which can be also identified through metabolomics. The common patterns of these biomarkers selected by LASSO across various cancers in our study, predominantly corresponded to dysregulated pathways in cancer, as reported in previous research, encompassing carbohydrate metabolism and amino acid metabolism (22, 25). Specifically, it exhibited outstanding performance by adeptly discerning cancers from healthy subjects, comparable to the performance of the model built by Zhang et al. (26), which incorporated blood-derived metabolites. Differently, our model utilized less invasive urine samples. Furthermore, besides effectively discerning between healthy individuals and cancer patients, we also exhibited robust discriminative capability between cancer patients and individuals with benign conditions, as these systemic factors have been documented to influence the metabolic profile (27).

During the early stages of tumor development, there is a demand for nutrient uptake and biosynthesis, leading to the metabolic reprogramming (28). Remarkably, in our investigation, we noted a decline in the model's sensitivity for detecting cancer at stage IV compared to the earlier stages, such as stages II and III in comparison with the non-cancerous group. We hypothesize that the irregular variations in selected metabolites across various stages of cancer progression could be the underlying factor, corroborated by prior research indicating the specific metabolic changes in tumor progression (28). For instance, MYC, a proto-oncogene, promotes the conversion of pyruvate to alanine, whereas in advanced malignant lesions, stimulates the transformation of pyruvate into lactate (29). Investigation of the metabolic evolution from preneoplasia to lung adenocarcinoma, researchers observed a gradual alteration in metabolic pathways, suggesting that cancer progression may not uniformly change metabolic patterns (30). In another study, metabolic profiles of patients with chronic obstructive pulmonary disease (COPD) were compared with those of lung cancer patients across various stages. The findings revealed notable differences in certain metabolite levels between early-stage lung cancer and COPD patients, while metabolite levels in late-stage lung cancer patients closely resembled those of COPD patients. This suggests that the serum metabolome may reflect lung dysfunction and systemic changes in advanced stages, which are absent in early stages (31). This finding may suggest the advantages of using metabolomics for early diagnosis of cancer, however, the specific mechanism needs to be further investigated.

In a cancer-screening scenario, a high specificity was uniformly predefined to diminish false positivity and thereby prevent overdiagnosis and anxiety. However, such specificity may compromise sensitivity to some extent, potentially undermining the benefits of multi-cancer early detection. For individuals with high-risk factors such as ulcerative colitis or adenomatous polyps for CRC (23), emphysema, chronic bronchitis, pneumonia, and tuberculosis for lung cancer (32), atrophic gastritis, and *H. pylori* infection for GC (33), a higher sensitivity is preferable, even with a slight compromise in specificity. In our study, the higher specificity of the model (> 99%) may be applicable for low-risk populations, which showcased higher diagnostic sensitivity compared to

conventional blood tumor marker-CEA, indicating its potential suitability for routine health check-up. The slightly lower specificity (95%) for discriminating cancer patients from non-cancerous groups (including HC and NCD) might be more suitable for high-risk populations to reduce the likelihood of missed cancer detection for clinical use. In practice, determining the balance between sensitivity and specificity requires further exploration through cost-effectiveness studies in various intended populations.

In our study, we selected metabolites that are commonly dysregulated across three different types of cancer and uniquely dysregulated within specific cancer subtypes. These metabolites were mapped in KEGG metabolic pathways, providing robust biomarkers for our model (34). Specifically, the features across cancer types were selected (LC vs. Non-LC, GC vs. CRC) to build the tumor classification model. Several studies have shown that gastrointestinal cancers shared similar mechanisms (24). Thus, we first compared the LC and gastrointestinal cancers (including GC and CRC), these selected biomarkers were mainly involved in amino acid metabolism. Furthermore, the biomarkers involved in carbohydrate metabolism were chosen to discriminate against GC and CRC. In a meta-analysis aimed at identifying biomarkers distinguishing between GC and CRC, it was observed that galactose metabolism was significantly enriched only in the GC group (24), consistent with our study findings. Additionally, utilizing TCGA pan-cancer database analysis, dysregulation of genes involved in fructose and mannose metabolism was observed specifically in GC patients, whereas no such dysregulation was found in the CRC group (35).

There are still several limitations in our study. The comprehensive follow-up was lacking for all non-cancerous participants to ensure their non-cancerous status, which could potentially result in an inflated false positive rate (FPR) and an underestimated positive predictive value (PPV). Besides, it's noteworthy that the variability in metabolites levels among both healthy individuals and those with diseases necessitates the validation of metabolic biomarkers using independent assay platforms and external validation cohorts. When millions of individuals undergo cancer screening, even a highly selective assay utilized in population screening may produce a significant number of false positives. Hence, it's crucial to eliminate false positives, as they could lead to significant and unnecessary stress for the individual. Furthermore, the current study only analyzed three cancer types, although these are among the most prevalent cancers in China, expanding the analysis to include other prevalent and significant cancers in the multi cancer strategy, such as liver, kidney, gynecological, and pancreatic cancers, would certainly broaden its clinical applicability.

Conclusions

In summary, we have developed a urine-based metabolic biomarker panel for early detection of multi-cancers. The selected metabolic panel, identified through the integration of algorithms and biological significance, effectively differentiated cancer patients from healthy controls and those with benign conditions. Additionally, a cancer classification model accurately predicted

tumor origin. Our findings demonstrated that urinary metabolomics can be applied as a universally applicable, straightforward, and cost-effective method for early cancer detection across large populations. In the future, with larger-scale external validation and the expansion to more types of cancer, the pan-cancer screening method based on urine metabolomics has the potential to be translated into clinical settings to support precision medicine, aiding in the identification and localization of common malignancies.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Institutional Review Board of the First Affiliated Hospital of Nanchang University (NCH; Nanchang, China), the Second Xiangya Hospital of Central South University (XYH; Changsha, China), and the Second Hospital of Tianjin Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

XX: Writing – review & editing, Project administration, Resources. CZ: Project administration, Resources, Writing – review & editing. BQ: Project administration, Resources, Writing – review & editing. YH: Conceptualization, Investigation, Writing – original draft. GS: Project administration, Resources, Writing – review & editing. JW: Investigation, Writing – original draft. SY: Investigation, Methodology, Writing – review & editing. TZ: Methodology, Writing – review & editing. QW: Resources, Writing – review & editing. LL: Conceptualization, Writing – review & editing. HW: Writing – review & editing. JH: Conceptualization, Writing – review & editing. WZ: Project administration, Resources, Writing – review & editing. YL:

Conceptualization, Writing – review & editing. YC: Project administration, Resources, Writing – review & editing. ZX: Project administration, Resources, Writing – review & editing.

Conflict of interest

YH, JW, SY, TZ, QW, LL, HW, JH, and YL are employees of Metanotitia Inc., Harbin.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1449103/full#supplementary-material

References

1. Serjomataram I, Bray F. Planning for tomorrow: global cancer incidence and the role of prevention 2020–2070. *Nat Rev Clin Oncol*. (2021) 18:663–72. doi: 10.1038/s41571-021-00514-z

2. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2022. *J Natl Cancer Cent*. (2024) 4:47–53. doi: 10.1016/j.jncc.2024.01.006

3. Hackshaw A, Clarke CA, Hartman AR. New genomic technologies for multi-cancer early detection: Rethinking the scope of cancer screening. *Cancer Cell*. (2022) 40:109–13. doi: 10.1016/j.ccell.2022.01.012

4. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. (2020) 31:745–59. doi: 10.1016/j.annonc.2020.02.011

5. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol*. (2021) 32:1167–77. doi: 10.1016/j.annonc.2021.05.806

6. Lennon AM, Buchanan AH, Kinde I, Warren A, Sheridan K, Rosica D, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science*. (2020) 369:eabb9601. doi: 10.1126/science.abb9601

7. Nguyen VTC, Nguyen TH, Doan NNT, Pham TMQ, Nguyen GTH, Nguyen TD, et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. *Elife*. (2023) 12:1–28. doi: 10.7554/ELIFE.89083

8. Choi H, Na KJ. Pan-cancer analysis of tumor metabolic landscape associated with genomic alterations. *Mol Cancer*. (2018) 17:1–6. doi: 10.1186/s12943-018-0895-9

9. Hocher B, Adamski J. Metabolomics for clinical use and research in chronic kidney disease. *Nat Rev Nephrol*. (2017) 13:269–84. doi: 10.1038/nrneph.2017.30

10. Bax C, Lotesoriere BJ, Sironi S, Capelli L. Review and comparison of cancer biomarker trends in urine as a basis for new diagnostic pathways. *Cancers (Basel)*. (2019) 11:1244. doi: 10.3390/cancers11091244

11. Woo HM, Kim KM, Choi MH, Jung BH, Lee J, Kong G, et al. Mass spectrometry based metabolomic approaches in urinary biomarker study of women's cancers. *Clin Chim Acta*. (2009) 400:63–9. doi: 10.1016/j.cca.2008.10.014

12. Zhang A, Sun H, Han Y, Yuan Y, Wang P, Song G, et al. Exploratory urinary metabolic biomarkers and pathways using UPLC-Q-TOF-HDMS coupled with pattern recognition approach. *Analyst*. (2012) 137:4200–8. doi: 10.1039/c2an35780a

13. Dinges SS, Hohm A, Vandergrift LA, Nowak J, Habbel P, Kaltashov IA, et al. Cancer metabolomic markers in urine: evidence, techniques and recommendations. *Nat Rev Urol*. (2019) 16:339–62. doi: 10.1038/s41585-019-0185-3

14. Mallafré-muro C, Llambrich M, Cumeras R, Pardo A, Brezmes J, Marco S, et al. Comprehensive volatilome and metabolome signatures of colorectal cancer in urine: A systematic review and meta-analysis. *Cancers (Basel)*. (2021) 13:2534. doi: 10.3390/cancers13112534

15. Mathé EA, Patterson AD, Haznadar M, Manna SK, Krausz KW, Bowman ED, et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res*. (2014) 74:3259–70. doi: 10.1158/0008-5472.CAN-14-0109

16. Chan AW, Mercier P, Schiller D, Bailey R, Robbins S, Eurich DT, et al. ^{1}H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *Br J Cancer*. (2016) 114:59–62. doi: 10.1038/bjc.2015.414

17. Shi Z, Li H, Zhang W, Chen Y, Zeng C, Kang X, et al. A comprehensive mass spectrometry-based workflow for clinical metabolomics cohort studies. *Metabolites*. (2022) 12:1168. doi: 10.3390/metabo12121168

18. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc*. (2006) 1:387–96. doi: 10.1038/nprot.2006.59

19. Chen Y, Shen G, Zhang R, He J, Zhang Y, Xu J, et al. Combination of injection volume calibration by creatinine and MS signals' normalization to overcome urine variability in LC-MS-based metabolomics studies. *Anal Chem*. (2013) 85:7659–65. doi: 10.1021/ac401400b

20. Shi Z, Wen B, Gao Q, Zhang B. Feature selection methods for protein biomarker discovery from proteomics or multiomics data. *Mol Cell Proteomics*. (2021) 20:100083. doi: 10.1016/J.MCPRO.2021.100083

21. Farshidfar F, Weljie AM, Kopciuk K, Buie WD, MacLean A, Dixon E, et al. Serum metabolomic profile as a means to distinguish stage of colorectal cancer. *Genome Med*. (2012) 4:1–13. doi: 10.1186/gm341

22. Platten M, Wick W, Van Den Eynde BJ. Tryptophan catabolism in cancer: Beyond IDO and tryptophan depletion. *Cancer Res*. (2012) 72:5435–40. doi: 10.1158/0008-5472.CAN-12-0569

23. Gao Q, Lin YP, Li BS, Wang GQ, Dong LQ, Shen BY, et al. Unintrusive multi-cancer detection by circulating cell-free DNA methylation sequencing (THUNDER): development and independent validation studies. *Ann Oncol*. (2023) 34:486–95. doi: 10.1016/j.annonc.2023.02.010

24. Tian J, Xue W, Yin H, Zhang N, Zhou J, Long Z, et al. Differential metabolic alterations and biomarkers between gastric cancer and colorectal cancer: A systematic review and meta-analysis. *Onco Targets Ther*. (2020) 13:6093–108. doi: 10.2147/OTT.S247393

25. Seth Nanda C, Venkateswaran SV, Patani N, Yuneva M. Defining a metabolic landscape of tumours: genome meets metabolism. *Br J Cancer*. (2020) 122:136–49. doi: 10.1038/s41416-019-0663-7

26. Zhang H, Zhao L, Jiang J, Zheng J, Yang L, Li Y, et al. Multiplexed nanomaterial-assisted laser desorption/ionization for pan-cancer diagnosis and classification. *Nat Commun*. (2022) 13:617. doi: 10.1038/s41467-021-26642-9

27. Patel S, Ahmed S. Emerging field of metabolomics: Big promise for cancer biomarker identification and drug discovery. *J Pharm BioMed Anal*. (2015) 107:63–74. doi: 10.1016/j.jpba.2014.12.020

28. Faubert B, Solmonson A, DeBerardinis RJ. Metabolic reprogramming and cancer progression. *Science*. (2020) 368:eaaw5473. doi: 10.1126/science.aaw5473

29. Hu S, Balakrishnan A, Bok RA, Anderton B, Larson PEZ, Nelson SJ, et al. ^{13}C-pyruvate imaging reveals alterations in glycolysis that precede c-Myc-induced tumor formation and regression. *Cell Metab*. (2011) 14:131–42. doi: 10.1016/j.cmet.2011.04.012

30. Nie M, Yao K, Zhu X, Chen N, Xiao N, Wang Y, et al. Evolutionary metabolic landscape from preneoplasia to invasive lung adenocarcinoma. *Nat Commun*. (2021) 12:1–13. doi: 10.1038/s41467-021-26685-y

31. Deja S, Porebska I, Kowal A, Zabek A, Barg W, Pawelczyk K, et al. Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. *J Pharm BioMed Anal*. (2014) 100:369–80. doi: 10.1016/j.jpba.2014.08.020

32. Marcus MW, Raji OY, Field JK. Lung cancer screening: Identifying the high risk cohort. *J Thorac Dis*. (2015) 7:S156–62. doi: 10.3978/j.issn.2072-1439.2015.04.19

33. Cai Q, Zhu C, Yuan Y, Feng Q, Feng Y, Hao Y, et al. Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: A nationwide multicentre study. *Gut*. (2019) 68:1576–87. doi: 10.1136/gutjnl-2018-317556

34. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Med*. (2016) 8:1–14. doi: 10.1186/s13073-016-0289-9

35. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. *Nat Commun*. (2018) 9:5330. doi: 10.1038/s41467-018-07