Establishment of a pattern recognition metabolomics model for the diagnosis of hepatocellular carcinoma

Peng-Cheng Zhou, Lun-Quan Sun, Li Shao, Lun-Zhao Yi, Ning Li, Xue-Gong Fan

ORCID number: Peng-Cheng Zhou 0000-0003-1536-8732; Lun-Quan Sun 0000-0002-0749-1995; Li Shao 0000-0001-8255-4362; Lun-Zhao Yi 0000-0002-1111-1510; Ning Li 0000-0002-4508-539X; Xue-Gong Fan 0000-0001-8081-348X.

Author contributions: Li N and Fan XG contributed to the experimental design and contributed equally to this work; Zhou PC collected the serum samples; Yi LZ performed the UPLC-MS analysis; Zhou PC, Sun LQ and Shao L contributed to the data analysis and wrote the original draft; all authors have read and approved the manuscript.

Supported by National Natural Science Foundation of China, No. 81800472 and No. 81670538; the Science Foundation of Hunan Health Commission, No. B2019184.

Institutional review board statement: The study was approved by the Ethics Committee of Xiangya Hospital, Central South University (Changsha, China).

Informed consent statement: The patients gave informed consent.

Conflict-of-interest statement: The authors have declared that no competing interests exist.

Data sharing statement: Technical

Abstract

BACKGROUND
Early diagnosis of hepatocellular carcinoma may help to ensure that patients have a chance for long-term survival; however, currently available biomarkers lack sensitivity and specificity.

AIM
To characterize the serum metabolome of hepatocellular carcinoma in order to develop a new metabolomics diagnostic model and identifying novel biomarkers for screening hepatocellular carcinoma based on the pattern recognition method.

METHODS
Ultra-performance liquid chromatography-mass spectroscopy was used to
characterize the serum metabolome of hepatocellular carcinoma (n = 30) and cirrhosis (n = 29) patients, followed by sequential feature selection combined with linear discriminant analysis to process the multivariate data.

RESULTS
The concentrations of most metabolites, including proline, were lower in patients with hepatocellular carcinoma, whereas the hydroxypurine levels were higher in these patients. As ordinary analysis models failed to discriminate hepatocellular carcinoma from cirrhosis, pattern recognition analysis was used to establish a pattern recognition model that included hydroxypurine and proline. The leave-one-out cross-validation accuracy and area under the receiver operating characteristic curve analysis were 95.00% and 0.90 [95% Confidence Interval (CI): 0.81-0.99] for the training set, respectively, and 78.95% and 0.84 (95%CI: 0.67-1.00) for the validation set, respectively. In contrast, for α-fetoprotein, the accuracy and area under the receiver operating characteristic curve were 65.00% and 0.69 (95%CI: 0.52-0.86) for the training set, respectively, and 68.42% and 0.68 (95%CI: 0.41-0.94) for the validation set, respectively. The Z test revealed that the area under the curve of the linear discriminant analysis model was significantly higher than the area under the curve of α-fetoprotein (P < 0.05) in both the training and validation sets.

CONCLUSION
Hydroxypurine and proline might be novel biomarkers for hepatocellular carcinoma, and this disease could be diagnosed by the metabolomics model based on pattern recognition.

Key words: Hepatocellular carcinoma; Pattern recognition; Metabolomics; Biomarkers

©The Author(s) 2020. Published by Baishideng Publishing Group Inc. All rights reserved.

Core tip: We used ultra-performance liquid chromatography-mass spectrometry to characterize the metabolome of serum samples from patients with hepatocellular carcinoma. We processed multivariate data using pattern recognition analysis and established a diagnostic model that included hydroxypurine and proline. The accuracy and area under the curve were 95.00% and 0.90 for the training set, respectively, and 78.95% and 0.84 for the validation set, respectively. The Z test revealed that the area under the curve of the model was significantly higher than that of α-fetoprotein. The results suggest that hydroxypurine and proline might be novel biomarkers for hepatocellular carcinoma, and the pattern recognition metabolomics model could be used to diagnose hepatocellular carcinoma.

INTRODUCTION
Hepatocellular carcinoma (HCC) is the fifth most common cancer and the third leading cause of death due to cancer worldwide[7,8]. In particular, approximately 50% of the total patients with HCC in the world are from China, owing to the highest carrier prevalence of hepatitis B[3-4]. Early diagnosis of HCC offers patients a better chance for long-term survival[5]. Although imaging technologies such as magnetic resonance imaging and ultrasonography, and serum biomarkers [notably α-fetoprotein (AFP)] are widely used to diagnose HCC in the clinic[6], they are far from satisfactory because they lack sensitivity and specificity[6]. Therefore, there is an urgent and unmet desire for novel screening methods and new biomarkers.

The emergence of metabolomics has provided a powerful tool for discovering novel biomarkers and revealing metabolic pathways of cancer and liver diseases[2,8]. A
metabolomics approach to screen individual metabolites or their combinations for the
diagnosis of HCC\textsuperscript{[10]} identified a series of potential biomarkers including phenylalanyl-
tryptophan, glycocholate, concanavamine succinic acid, bile acid, long chain fatty acid,
and so on for future clinical application\textsuperscript{[5,7,11]. However, none of these markers have
thus far been validated for clinical applications. Metabolomics datasets commonly
contain hundreds to thousands of variables; however, biomarkers are identified using
conventional data processing methods such as principal component analysis (PCA),
partial least squares discriminant analysis (PLS-DA), orthogonal partial least squares
discriminant analysis (OPLS-DA), and binary logistic regression\textsuperscript{[11,12]. With the advent
of data processing technology to handle big data, it is incumbent upon researchers in
this area to adopt advanced methods such as pattern recognition to seek new
biomarkers and to establish mathematical models that facilitate screening for HCC.

In previous studies, we established a pattern recognition metabolomics method
based on sequential feature selection combined with linear discriminant analysis
(LDA) to evaluate the severity of fulminant hepatic failure and for the differential
diagnosis of \textit{Clostridium difficile} infection\textsuperscript{[13,14]. In the current study, ultra-performance
liquid chromatography-mass spectroscopy (UPLC-MS) was used to characterize the
serum metabolomes of patients with HCC, patients with cirrhosis, and healthy
controls. Furthermore, the pattern recognition method developed herein was used to
process multivariate data with the aim of developing a novel metabolomics diagnostic
model and identifying novel biomarkers for HCC screening purposes.

**MATERIALS AND METHODS**

**Patients and samples**
Between March and August 2016, samples from patients who met the inclusion criteria
of HCC diagnosis set by the Ministry of Health were collected\textsuperscript{[15]. HCC confirmation
required histological evidence or two different imaging techniques, or the combination
of one imaging technique and an AFP level of > 400 ng/mL. Patients with cirrhosis
meeting the criteria described elsewhere\textsuperscript{[16] based on clinical manifestations, laboratory
examinations, and imaging results were included. HCC patients (C group, \(n = 30\)) all
had cirrhosis, and cirrhosis patients without HCC were included in Y group (\(n = 29\)).
The Child-Pugh Score in patients in the C group and Y group patients was A or B. Healthy controls (N group, \(n = 31\)) were chosen from the general population. The
exclusion criteria were Child-Pugh Score C patients, malignant neoplasm (except HCC
for C group), metabolic diseases, autoimmune disease, excess alcohol consumption,
and known history of toxic exposure. Whole blood samples (3-5 mL) were collected on
an empty stomach in the morning in BD \textsuperscript{®} blood specimen collection tubes
(Weigao Group, Weihai, China). Whole blood samples were stored at 4°C immediately
after collection and were transported to the laboratory in < 30 min. After
centrifugation at 3000 \(\times g\) for 10 min at 4°C, a portion of the serum from the samples
was used for biochemical assays and the remaining serum was aliquoted into fresh
Eppendorf \textsuperscript{®} tubes and stored at -80°C for metabolomic analysis. Fresh surgical tumor
tissue samples were obtained from patients following informed consent.

**Virology, biochemical parameters, and histopathology assay**
Hepatitis B virus (HBV) and HCV antigens and a biochemical panel including alanine
aminotransferase, aspartate aminotransferase, glutamic-oxaloacetic transaminase, total
bilirubin, direct bilirubin, total protein, and albumin were assayed in the clinical
laboratory. Histopathological samples were prepared as described previously\textsuperscript{[13].

**Chemicals and reagents**
Acetonitrile and methanol (HPLC grade) were purchased from Merck (Darmstadt,
Germany). Distilled water was purified using a Milli-Q system (Darmstadt, Germany).
Fatty acids, amino acids, bile acid, and nucleotide standards were purchased from Sigma-Aldrich (St. Louis, MO, United States). Citric acid, pantothenic acid, and
malonic acid were purchased from Supelco (Bellefonte, PA, United States).
Lyosphosphatidyl cholines (LysoPCs) and lyosphosphatidyl ethanolamine were
purchased from Avanti Polar Lipids, Inc. (AL, United States).

**Sample preparation**
Prior to the assay, all samples were thawed on ice. Pooled aliquots (1 \(\mu\)L) of each
sample formed the quality control (QC) sample. Metabolites in serum were extracted
by methanol (serum/methanol (V/V) = 1:3). The mixture (100 μL) was vortexed for 60 s, and then centrifuged at 14000 × g for 10 min at 4°C. Supernatants were dried by nitrogen flow and then re-dissolved in 100 μL methanol. The mixture was again centrifuged at 14000 × g for 5 min at 4°C. The resulting clear supernatant was transferred into UPLC vials and stored at 4°C.

**UPLC-MS assay**

An aliquot (2 μL) of the clear supernatant obtained above was chromatographed on a Thermo Fisher Scientific UltiMate 3000 UPLC system using an ACQUITY UPLC BEH C18 analytical column (i.d. 2.1 mm × 100 mm, particle size 1.7 mm, pore size 130 Å). Mobile phase A and mobile phase B were water/formic acid (99.9: 0.1, V/V) and acetonitrile/formic acid (99.9: 0.1, V/V), respectively, and the flow rate was 200 μL/min. A linear gradient was optimized as follows: the initial composition of the mobile phase was 95% A and 5% B; 0-2 min, 95% A; 2-9 min, 95%-62% A; 9-14 min, 62%–32% A; 14-22 min, 32%-0% A; 22-30 min, 0-95% A. The column eluent was directed to the mass spectrometer for analyses.

Mass spectrometry was performed on a Thermo Fisher Scientific Q-Exactive Focus Mass Spectrometer operating in positive ion electrospray mode. The instrument parameters were set as follows: Mass range scanned from 50 to 1000, spray voltage was 4000 V, atomization temperature was 300°C, nebulizer pressure was 45 bar, capillary temperature was 350°C, and the capillary voltage was set to 4.00 kV; the sampling cone voltage was set to 35.0 V. The instrument parameters for MS/MS analysis were set at different collision energies according to the stability of metabolites (collision energy was set from 15 to 35 eV).

Five injections of QC samples were performed to equilibrate the UPLC-MS systems prior to testing individual patient samples. QC samples were injected after every six patient samples at regular intervals throughout the analytical run. Patient samples were tested in a random manner.

**Data processing and statistical analysis**

The raw UPLC-MS data of the samples were extracted using MZmine2.3 software and Xcalibur software (Thermo Fisher Scientific), which enabled detection, integration and normalization of the intensities of the peaks to the sum of peaks within the sample and to create a multivariate dataset containing the retention time, m/z, and relative abundances. The parameters were set as follows: Retention time ranging from 0 to 30 min, mass range m/z from 50 to 1000, and mass tolerance at 0.05 Da. For peak integration, peak width at 5% of the height was 1 s, peak-to-peak baseline noise was 0, peak intensity threshold was 100, and retention time window was 0.20 s.

The statistical analysis is shown in **Figure 1**. In brief, we used SIMCA-P + 12.0 software (Umetrics, AB, Sweden) to perform PCA, PLS-DA, and OPLS-DA. Pattern recognition analysis based on sequential feature selection combined with LDA for diagnosis of HCC, and the Z test [for comparison of area under curve (AUC)] were performed using Matlab Version 8.1 (R2013a) software (MathWorks Inc., Natick, MA, United States). One-way ANOVA, the Chi-square test, and Kruskal–Wallis test were conducted using SPSS v16.0 software (SPSS Inc. Chicago, IL, United States). Differences were considered statistically significant at \( P < 0.05 \).

**Marker identification**

The compounds were identified by searching the Human Metabolome Database (http://hmdb.ca/), PubChem compound database (http://www.ncbi.nlm.nih.gov), and our own compound database that includes metabolites previously identified by us. Finally, the compound was verified by comparing the mass spectra and retention time of potential biomarkers with authentic standards (Supplementary Figures 1-5).

**RESULTS**

**Study population and clinical characteristics**

Demographic data and clinical characteristics of the subjects are shown in **Table 1**. Thirty patients with HCC (all with cirrhosis, C group), 29 patients with cirrhosis (all without HCC, Y group), and 31 healthy controls (N group) were enrolled. There were no significant differences in age and sex among the three groups, and no significant differences in the causes of liver injury and Child-Pugh Score between C group and Y group. The levels of AFP, glutamic-oxaloacetic transaminase, and alanine...
Table 1 General characteristics of patients and healthy controls

| Characteristics       | C (n = 30) | Y (n = 29) | N (n = 31) | P value |
|-----------------------|------------|------------|------------|---------|
| Sex (Male/Female)     | 25/5       | 21/8       | 25/6       | 0.565   |
| Age (yr)              | 52.93 ± 11.01 | 56.63 ± 9.15 | 51.23 ± 11.79 | 0.148   |
| Pathogens             |            |            |            |         |
| HBV                   | 25         | 24         | /          | 0.720   |
| HCV                   | 1          | 2          |            |         |
| HBV + HCV             | 1          | 0          |            |         |
| None                  | 3          | 3          |            |         |
| AFP (ng/mL)           |            |            |            |         |
| > 200                 | 11         | 0          | /          | 0.000   |
| 50-199                | 4          | 1          |            |         |
| < 50                  | 15         | 28         |            |         |
| ALT (U/L)             | 162.32 ± 201.06 | 91.02 ± 156.39 | 20.34 ± 8.43 | 0.000   |
| AST (U/L)             | 146.35 ± 112.70 | 114.49 ± 191.67 | 21.59 ± 4.51 | 0.012   |
| TBIL (μmol/L)         | 39.21 ± 68.38 | 40.87 ± 42.41 | 9.66 ± 2.66 | 0.015   |
| DBIL (μmol/L)         | 17.91 ± 34.43 | 17.90 ± 23.03 | 4.49 ± 1.38 | 0.044   |
| TP (g/L)              | 62.42 ± 10.95 | 74.14 ± 8.05 | 72.31 ± 3.96 | 0.000   |
| ALB (g/L)             | 33.51 ± 6.30 | 37.65 ± 7.64 | 45.36 ± 2.62 | 0.000   |
| Child-Pugh score (A/B)| 18/12      | 15/14      | /          | 0.353   |

AFP: α-fetoprotein; ALB: Albumin; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; DBIL: Direct bilirubin; TBIL: Total bilirubin; TP: Total protein.

Aminotransferase were relatively higher and the level of albumin was relatively lower in patients with HCC than in patients with cirrhosis and healthy controls. The histopathology results of patients with HCC are shown in Supplementary Figure 6. We used the Chinese staging system to stage HCC[15], and 11 cases were stage IIIa, 12 cases were stagellb, one case was stageIia, 5 cases were stagelb, and one case was stagela.

Quality control of UPLC-MS assay
QC samples clustered compactly in the middle of the PCA score plot (Figure 2A). The coefficient of variation (CV) of identified metabolites in QC samples ranged from 2.09% to 16.27% with a median CV of 7.83% (Table 2).

Metabolic profiles of serum samples
Patients with HCC, patients with cirrhosis, and healthy controls showed no significant differences in the base peak intensity chromatogram (Supplementary Figure 7). The three groups intermixed with each other in the PCA score plot, although there was a tendency to separate along PC1 (Figure 2B). Characterization of metabolic differences among the three groups using PLS-DA and OPLS-DA showed that the three groups also intermixed with each other in the PLS-DA score plot (Supplementary Figure 8). The PLS-DA score plot of the HCC group vs the cirrhosis group also intermixed with each other (Supplementary Figure 9). Validation plots of the PLS-DA models acquired through 20 permutation tests were used for cross-validation purposes (Supplementary Figures 10 and 11). Analysis of the PLS-DA score plot for all three groups revealed that R2 = (0.0, 0.401) and Q2 = (0.0, -0.35); cross-validation of the PLS-DA score plot of C group and Y group revealed that R2 = (0.0, 0.645) and Q2 = (0.0, -0.507). Although the PLS-DA model showed intermixing of the three groups, they could be separated in the OPLS-DA model (Figure 3A). OPLS-DA score plots of the HCC group vs healthy controls (Figure 3B), the cirrhosis group vs healthy controls (Figure 3C), and the HCC group vs the cirrhosis group (Figure 3D) demonstrated very clear separation. However, the R2 and Q2 values were not high enough in the three OPLS-DA models.

Biomarkers for HCC
Potential biomarkers were characterized by variable importance in the projection values retrieved from the PLS-DA model combined with the Kruskal–Wallis test (P <
### Table 2: Significantly altered metabolites

| Retention time (min) | m/z     | Metabolites          | Adduction | Adduct mass | Delta ppm | Coefficient of variation (%) | Comparison | C vs N | Y vs N | C vs Y |
|---------------------|---------|----------------------|-----------|-------------|-----------|-------------------------------|------------|--------|--------|--------|
| 9.57                | 166.0862| Phenylalanine        | M + H     | 166.0863    | 1.00      | 8.36                          | D          | U      | NS     |
| 3.49                | 118.0864| Valine               | M + H     | 118.0863    | 1.00      | 6.76                          | D          | NS     | NS     |
| 6.63                | 132.1019| Leucine              | M + H     | 132.1019    | 0.00      | 3.34                          | NS         | U      | NS     |
| 3.58                | 116.0708| Prolin               | M + H     | 116.0706    | 1.00      | 14.23                         | D          | D      | D      |
| 5.42                | 182.0811| Tyrosine             | M + H     | 182.0812    | 0.00      | 6.12                          | NS         | U      | NS     |
| 6.05                | 132.1019| Isoleucine           | M + H     | 132.1019    | 0.00      | 3.37                          | NS         | U      | D      |
| 4.89                | 150.0583| Methionine           | M + H     | 150.0583    | 0.00      | 9.28                          | NS         | U      | NS     |
| 3.16                | 156.0766| Histidine            | M + H     | 156.0768    | 1.00      | 10.83                         | D          | U      | D      |
| 3.44                | 148.0602| Glutamic acid        | M + H     | 148.0604    | 2.00      | 6.38                          | U          | D      | U      |
| 3.32                | 106.0502| Serine               | M + H     | 106.0499    | 3.00      | 2.59                          | U          | D      | U      |
| 3.38                | 147.0762| Glutamine            | M + H     | 147.0764    | 1.00      | 11.41                         | NS         | D      | D      |
| 3.44                | 90.0555 | Alanine              | M + H     | 90.0550     | 5.00      | 12.28                         | D          | D      | NS     |
| 5.43                | 165.0546| Hydroxycinnamic acid | M + H     | 165.0546    | 0.00      | 16.17                         | D          | NS     | D      |
| 5.43                | 123.0442| Benzolic acid        | M + H     | 123.0441    | 1.00      | 10.11                         | D          | U      | NS     |
| 9.57                | 149.0596| Cinnamic acid        | M + H     | 149.0597    | 1.00      | 12.29                         | D          | U      | NS     |
| 24.40               | 190.0497| Kynurenic acid       | M + H     | 190.0499    | 1.00      | 6.51                          | D          | D      | U      |
| 26.39               | 169.0495| Vanillic acid        | M + H     | 169.0495    | 0.00      | 3.41                          | D          | D      | U      |
| 13.83               | 239.0912| Trimethoxycinnamic acid | M + H     | 239.0914    | 1.00      | 5.08                          | D          | U      | NS     |
| 18.85               | 279.2318| Linoleic acid        | M + H     | 279.2319    | 0.00      | 11.26                         | D          | NS     | D      |
| 3.10                | 130.0862| Pipecolinic acid     | M + H     | 130.0863    | 0.00      | 10.58                         | D          | U      | NS     |
| 29.42               | 494.3235| LysoPC 16:1          | M + H     | 494.3241    | 1.00      | 4.32                          | NS         | D      | D      |
| 22.87               | 542.3234| LysoPC 20:5          | M + H     | 542.3241    | 1.00      | 3.09                          | NS         | D      | NS     |
| 17.33               | 548.3705| LysoPC 20:2          | M + H     | 548.3711    | 1.00      | 2.31                          | D          | D      | NS     |
| 21.65               | 550.3857| LysoPC 20:1          | M + H     | 550.3867    | 2.00      | 5.58                          | D          | D      | NS     |
| 23.13               | 468.3078| LysoPC 14:0          | M + H     | 468.3085    | 1.00      | 6.27                          | D          | D      | D      |
| 19.25               | 478.2926| LysoPE 20:1          | M + H     | 478.2928    | 0.00      | 8.72                          | NS         | D      | NS     |
| 17.58               | 181.0857| Propylparaben        | M + H     | 181.0859    | 1.00      | 6.39                          | D          | NS     | NS     |
| 5.42                | 136.0756| Acetyllarylamine     | M + H     | 136.0757    | 0.00      | 2.59                          | D          | U      | D      |
| 18.20               | 127.0390| Trihydroxybenzene    | M + H     | 127.0390    | 0.00      | 13.83                         | D          | U      | D      |
| 22.22               | 191.1428| Damascenone          | M + H     | 191.1430    | 1.00      | 10.02                         | U          | D      | NS     |
| 10.70               | 181.0718| Myoinositol         | M + H     | 181.0707    | 6.00      | 8.55                          | NS         | D      | NS     |
| 4.88                | 137.0457| Hydroxyprurine       | M + H     | 137.0458    | 1.00      | 9.74                          | U          | D      | U      |
| 3.48                | 114.0664| Creatinine           | M + H     | 114.0662    | 2.00      | 7.83                          | D          | NS     | NS     |
| 3.82                | 72.0815 | Pyrrolidine          | M + H     | 72.0808     | 10.00     | 2.09                          | U          | U      | NS     |
| 11.71               | 195.0873| Methyl lucopyranoside| M + H     | 195.0863    | 6.00      | 12.43                         | D          | NS     | NS     |

LysoPC: Lysophosphatidyl choline; LysoPE: Lysophosphatidyl ethanolamine; U: Upregulated; D: Decreased; NS: No statistical difference.

Potential biomarkers were identified by a preliminary search of the HMDB and
Figure 1  Road map of data analysis. Road map of data analysis. Ordinary multivariate statistical analysis (principal component analysis, partial least squares discriminant analysis, and orthogonal partial least squares discriminant analysis) were used to describe the metabolome of the three groups. Pattern recognition analysis based on sequential feature selection combined with linear discriminant analysis were used to diagnose hepatocellular carcinoma. The Kruskal–Wallis test was used to identify differences in metabolites. PCA: Principal component analysis; PLS-DA: Partial least squares discriminant analysis; OPLS-DA: Orthogonal partial least squares discriminant analysis; LDA: Linear discriminant analysis; HCC: Hepatocellular carcinoma.

PubChem compound databases and verified by comparing the mass spectra and retention time of potential biomarkers with authentic standards. As shown in Table 2 and Supplementary Figure 12, the levels of most metabolites, including proline, were lower in patients with HCC than in healthy controls and patients with cirrhosis (Figure 4A). However, the levels of glutamic acid, pyrrolidine, and damascenone were higher in patients with HCC than in healthy controls; glutamic acid, kynurenic acid, vanillic acid, and hydroxypurine (Figure 4B) were higher in patients with HCC than in patients with cirrhosis.

Pattern recognition for diagnosis of HCC

We intended to establish a PLS-DA model or OPLS-DA model with the aim of distinguishing patients with HCC from patients with cirrhosis. However, as the metabolomes of HCC and cirrhosis are not very different, the efficiency of the models was not robust enough to discriminate the two groups using ordinary PLS-DA or OPLS-DA models. Therefore, we used pattern recognition, an advance data processing method, to achieve our aim. To enable this, the dataset was randomly split into a training set and a validation set. The training set comprised 20 HCC samples and 20 cirrhosis samples, and the validation set comprised 10 HCC samples and nine cirrhosis samples. We used sequential feature selection to select the most suitable metabolites for constructing the best performing LDA model based on the training set. The validation set was used to confirm the reliability of the model for discriminating...
Figure 2  Principal component analysis. A: The principal component analysis score plot of all samples including quality control samples. R²X = 0.134 cum, Q² = 0.106 cum; and B: The principal component analysis score plot of all three groups, hepatocellular carcinoma group (C group), cirrhosis group (Y group), and healthy controls (N group). R²X = 0.139 cum, Q² = 0.103 cum. QC: Quality control; PCA: Principal component analysis; HCC: Hepatocellular carcinoma.

patients with HCC from patients with cirrhosis. When the metabolites hydroxypurine and proline were included in the LDA model, a differential distribution pattern between HCC and cirrhosis began to emerge in the LDA plot (Figure 5). The leave-one-out cross-validation analysis provided accuracy, sensitivity, specificity, a positive predictive value, and a negative predictive value of 95.00%, 100.00%, 90.00%, 0.91, and 1.00, respectively, for the training set, and 78.95%, 100.00%, 60.00%, 0.69, and 1.00, respectively, for the external validation set (Table 3). Validation of AFP as a biomarker to discriminate HCC and cirrhosis provided accuracy, sensitivity, specificity, a positive predictive value, and a negative predictive value of 65.00%, 30.00%, 100.00%, 1.00 and 0.59, respectively, for training samples, and 68.42%, 40.00%, 100.00%, 1.00 and 0.60, respectively, for test samples. For the training samples, the AUC in the LDA model (AUC_{LDA}) was 0.90 (95% CI: 0.81–0.99, P < 0.05, Figure 6A), and AUC_{AFP} was 0.69.
LDA: Linear discriminant analysis; ROC: Receiver operating characteristic curve; AUC: Area under curve.

(95%CI: 0.52–0.86, P < 0.05, Supplementary Figure 13); AUC_{LDA} was significantly more than AUC_{AFP} (P < 0.05, Z test). For validation samples, AUC_{LDA} was 0.84 (95%CI: 0.67–1.00, P < 0.05, Figure 6B), and AUC_{AFP} was 0.68 (95%CI: 0.41–0.94, P = 0.191, Supplementary Figure 14); AUC_{LDA} was significantly larger than AUC_{AFP} (P < 0.05, Z test).

**DISCUSSION**

In this study, the serum metabolomes of patients with HCC, patients with cirrhosis, and healthy controls were profiled by UPLC-MS to establish a metabolomics model for the diagnosis of HCC. This approach not only enabled elucidation of HCC pathogenesis but also provided a mathematical model based on possible biomarkers for screening HCC.

The stability of metabolomics data and the comparability of demographic data are the two crucial issues that should be considered prior to statistical analysis\(^9\). In this study, the reproducibility and stability of metabolomics data are reflected in the compact clustering of QC samples in the PCA score plot, as well as in the low CV of specific metabolites of the QC samples. There were no statistical differences in age and sex among the patients with HCC, patients with cirrhosis, and healthy controls. Also, the constituent ratio of etiology of liver injury (pathogenesis) was comparable between the HCC and cirrhosis groups, all of which confirm the reliability of the UPLC-MS assay and optimal homogeneity of baseline characteristics\(^9\).

The liver is the principal organ for metabolism of carbohydrates, lipids, amino acids *etc*\(^9\). Particularly in HCC, liver disease always results in apparent metabolic dysregulation\(^9\), as in the case of glutamine addiction, a hallmark feature of HCC\(^9\). The decrease in serum metabolites in patients with HCC is largely due to uptake and utilization of metabolites by the tumor to feed its malignant behavior, as in the case of glutamine addiction\(^9\). This is evident in HCC tissue that has 20 times higher glutaminase 1 concentration than normal liver tissue\(^9\), leading to 10 times faster consumption of glutamine resulting in diminished glutamine levels in the serum of patients with HCC. On the contrary, an increase in the concentration of serum metabolites in HCC may reflect tumor necrosis. The best illustration of this process is the increase in hydroxyprurine in the serum of patients with HCC, likely due to the release of nucleic acids from tumor tissues, which then metabolizes into hydroxyprurine under necrotic conditions\(^9\).

Our findings are in line with previous studies that demonstrated diminished levels of serum phospholipid metabolites in patients with liver diseases (including HCC, liver cirrhosis, hepatitis, and liver failure)\(^7\). Indeed, through an untargeted metabolomics approach, we found significantly reduced amounts of phospholipid metabolites in patients with HCC. Reduced serum LysoPC, a molecule associated with malignancies, autoimmune disease, inflammation, and cell signaling\(^17\), is an indicator of liver injury; LysoPC correlates with model for end-stage liver disease score, independently of age, sex, and diet. As the patients with HCC in our cohort also had concurrent liver cirrhosis, the serum LysoPC of C group was lower than that of healthy controls. However, since the severity of liver injury was similar between C and Y groups, the serum LysoPC concentration was not significantly different between these groups. Low levels of LysoPC may be attributed to the inhibition of phospholipase A2 or LCAT activity or perturbed LysoPC acyltransferase activity\(^9\). More recently, based on studies from our group and others, it was postulated that...
Figure 3  Metabolic profiles of serum from hepatocellular carcinoma patients, cirrhosis patients and healthy controls. A: The orthogonal partial least squares discriminant analysis (OPLS-DA) score plot for all the three groups. Model efficiency: R2X = 0.370 cum, R2Y = 0.838 cum, Q2 = 0.467 cum; B: The OPLS-DA score plot of C group and N group. R2X = 0.187 cum, R2Y = 0.790 cum, Q2 = 0.603 cum; C: The OPLS-DA score plot of Y group and N group. R2X = 0.559 cum, R2Y = 0.962 cum, Q2 = 0.696 cum; and D: The OPLS-DA score plot of C group and Y group. R2X = 0.274 cum, R2Y = 0.812 cum, Q2 = 0.358 cum. OPLS-DA: Orthogonal partial least squares discriminant analysis.

excessive consumption of LysoPC results in an anti-inflammatory response, leading to
The reduced levels of serum creatinine found in patients with HCC in this study may be attributed to the diminished hepatic conversion of creatine to creatinine in patients with hepatic disease\(^9\). Another reason may be the decrease in levels of serine and alanine, involved in the synthesis of creatine, in HCC\(^5\). Down regulation of fatty acids was also found in patients with HCC compared with cirrhotic patients and healthy controls. Fatty acids can be transported into the mitochondria for beta-oxidation to generate adenosine triphosphate (ATP) energy, and its metabolism could be perturbed in patients with chronic liver disease\(^2\). Thus, we hypothesized that differential levels of metabolites in HCC may enable biomarker identification for the diagnosis of HCC.

As the PCA and PLS-DA models suffered from relatively poor efficiencies in our study and were overfit for the dataset, they were therefore unable to discriminate patients with HCC from patients with cirrhosis. Hence, a pattern recognition approach, based on sequential feature selection combined with LDA, was adopted to find the most suitable combination of biomarkers. This resulted in the generation of an LDA model for the diagnosis of HCC, which included two novel biomarkers, hydroxypurine and proline, highlighting the rapid growth and necrotic characteristics of HCC. As the accuracy, sensitivity, negative predictive value, and AUC\(\text{LDA}\) were higher in the LDA model compared to those in the AFP diagnostic model, the...
Figure 5 Pattern recognition for the diagnosis of hepatocellular carcinoma. Pattern recognition analysis based on sequential feature selection combined with linear discriminant analysis (LDA) was used to find the most suitable biomarkers for discriminating hepatocellular carcinoma patients from cirrhosis patients in the training set. The validation set was used to confirm the reliability of the model. Hydroxypurine and proline were included in the LDA model. Function 1 and function 2 are the first two eigenvectors. Hepatocellular carcinoma samples and cirrhosis samples demonstrated different distributions in the LDA plot.

In conclusion, competitive access to nutrition and necrosis can be identified in HCC using a metabolomics model based on sequential feature selection combined with LDA, which may be an ideal method for novel biomarker discovery.
Figure 6  Receiver operating characteristic curve of the pattern recognition diagnostic model. A: Receiver operating characteristic curve for the training set of the linear discriminant analysis model. Area under the curve for the training set was 0.90 (95%CI: 0.81-0.99); B: Receiver operating characteristic for the validation (test) set of the linear discriminant analysis model. Area under the curve for the validation set was 0.84 (95%CI: 0.67-1.00).
ARTICLE HIGHLIGHTS

Research background
Early diagnosis of hepatocellular carcinoma (HCC) offers patients a better chance for long-term survival. The current biomarkers are far from satisfactory as they lack sensitivity and specificity. The emergence of metabolomics has provided a powerful tool for discovering novel biomarkers. In previous studies, we established a pattern recognition metabolomics method based on sequential feature selection combined with linear discriminant analysis for differential diagnosis.

Research motivation
There is an urgent and unmet desire for novel screening methods and new biomarkers for the diagnosis of HCC. Whether the pattern recognition method mentioned above could be used to establish a metabolomics model for the diagnosis of HCC is still unknown.

Research objectives
We aimed to use the pattern recognition method to develop a metabolomics diagnostic model and identify new biomarkers for HCC screening.

Research methods
We used ultra-performance liquid chromatography-mass spectroscopy to characterize the serum metabolome of HCC and cirrhosis patients. We then processed the multivariate data using sequential feature selection combined with linear discriminant analysis.

Research results
The concentrations of most metabolites, including proline, were lower in patients with HCC, whereas hydroxypurine levels were higher in these patients. As ordinary analysis models failed to discriminate hepatocellular carcinoma from cirrhosis, pattern recognition analysis was used to establish a pattern recognition model that included hydroxypurine and proline. The leave-one-out cross-validation accuracy and area under curve (AUC) were 95.00% and 0.90 (95% confidence interval (CI): 0.81–0.99) for the training set, respectively, and 78.95% and 0.84 (95%CI: 0.67–1.00) for the validation set, respectively. The Z test revealed that the AUC of the model was significantly higher than the AUC ($P < 0.05$) in both the training and validation sets.

Research conclusions
Hydroxypurine and proline might be novel biomarkers for HCC, and the disease could be diagnosed by the metabolomics model based on pattern recognition.

Research perspectives
This study determined the applicability of the pattern recognition metabolomics model for the diagnosis of HCC. Two novel biomarkers for HCC were also found. Future studies should verify the validity of the model and the applicability of the biomarkers in the early diagnosis of patients with HCC.

REFERENCES

1. Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, MacIntyre MF, Allen C, Hansen G, Woodbrook R, Wolfe C, Hamadeh RR, Moore A, Werdecker A, Gessner BD, Te Ao B, McMahon B, Karimkhani C, Yu C, Cooke GS, Schwebel DC, Carpenter DO, Pereira DM, Nash D, Kazi DS, De Leo D, Plass D, Ukwaja KN, Thurston GD, Yun Jin K, Simard EP, Mills E, Park EK, Catalá-López F, deVeber G, Gotay C, Khan G, Hsogood HD 3rd, Santos IS, Leasher JL, Singh J, Leigh J, Jonas JB, Sanabria J, Beardsley J, Jacobson KH, Takahashi K, Franklin RC, Ronfani L, Montico M, Naldi L, Tonelli M, Geleijnse J, Petzold M, Shimie MG, Younis M, Yonemoto N, Breitborde N, Vip P, Pourmalek F, Lotufo PA, Estehghamati A, Hankey GJ, Ali R, Lunevicius R, Malekzadeh R, Delavalle R, Weintrub R, Lucas R, Hay R, Rojas-Rueda D, Westerman R, Sepanlou SG, Nolte S, Patten S, Weichenthal S, Abere SF, Ferestehnejad SM, Shiue I, Driscoll T, Vasankari T, Afsarif U, Rahimi-Movaghar V, Vlassov VV, Marceres WS, Mekonnen W, Melaku YA, Yano Y, Artaman A, Campos I, MacLachlan J, Mueller U, Kim D, Trillini M, Esthagi B, Williams HC, Shibuya K, Dandona R, Murthy K, Cowie B, Amare AT, Antonio CA, Castañeda-Orjuela C, van Gool CH, Violante F, Ol IH, Deribe K, Soreide K, Knibbs L, Kereselidze M, Green M, Cardenas R, Roy N, Tillmann T, Li Y, Krueger H, Monasta L, Dey S, Sheikhbahaei S, Hafezi-Nejad N, Kumar GA, Sreeramareddy CT, Dandona L, Wang H, Vollset SE, Mokdad
Xiao JF, Varghese RS, Zhou B, Nezami Ranjbar MR, Zhao Y, Tsai TH, Di Poto C, Wang J, Goerlitz D, Luo P, Bie P, Liu LX, Wen TF, Han GH, Jia P, Liu P, Jia Y, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu QX, He J. Cancer statistics in China, 2016; CA Cancer J Clin 2016; 66: 115-132 [PMID: 26808342 DOI: 10.3322/caac.21338]

Fu S, Li N, Zhou PC, Huang Y, Zhou RR, Fan XG. Detection of HBV DNA and antigens in HBsAg-positive patients with primary hepatocellular carcinoma. Clin Res Hepatol Gastroenterol 2017; 41: 415-423 [PMID: 28286056 DOI: 10.1016/j.clinre.2017.01.009]

Xiao Y, Sun L, Fu Y, Huang Y, Zhou R, Hu X, Zhou P, Quan J, Li N, Fan XG. High mobility group box 1 promotes sorafenib resistance in HepG2 cells and in vivo. BMC Cancer 2017; 17: 857 [PMID: 29246127 DOI: 10.1186/s12885-017-3868-2]

Cheng T, Xie G, Wang X, Fan J, Quy Y, Zheng X, Qi X, Cao Y, Su M, Wang X, Xu LX, Yan Y, Liu P, Jia W. Serum and urine metabolite profiling reveals potential biomarkers of human hepatocellular carcinoma. Mol Cell Proteomics 2011; 10: M110.004495 [PMID: 21518826 DOI: 10.1074/mcp.M110.004495]

Ren B, Zou G, Xu F, Huang Y, Xu G, He J, Li Y, Zhu H, Yu P. Serum levels of anti-sperm-associated antigen 9 antibody are elevated in patients with hepatocellular carcinoma. Oncol Lett 2014; 7: 1600-1603 [PMID: 24934428 DOI: 10.3892/ol.2014.3752]

Wang B, Chen D, Chen Y, Hu Z, Cao M, Xie Q, Chen Y, Xu J, Zheng S, Li Z. Metabonomic profiles discriminate hepatocellular carcinoma from liver cirrhosis by ultra-performance liquid chromatography-mass spectrometry. J Proteome Res 2012; 11: 1217-1227 [PMID: 22200553 DOI: 10.1021/pr2006922]

Peng F, Liu Y, He C, Kong Y, Quyang X, Xie L, Liu Z, Peng J. Prediction of platinum-based chemotherapy efficacy in lung cancer based on LC-MS metabolomics approach. J Pharm Biomed Anal 2018; 154: 95-101 [PMID: 29544107 DOI: 10.1016/j.jpba.2018.02.051]

Zhau P, Shao L, Zhao L, Lv G, Pan X, Zhang A, Li J, Zhou N, Chen D, Li L. Efficacy of Fluidized Bed Bioartificial Liver in Treating Fulminant Hepatic Failure in Pigs: A Metabolomics Study. Sci Rep 2016; 6: 26070 [PMID: 27194381 DOI: 10.1038/srep26070]

Wang X, Zhang A, Sun H. Power of metabolomics in diagnosis and biomarker discovery of hepatocellular carcinoma. Hepatology 2013; 57: 2072-2077 [PMID: 23150189 DOI: 10.1002/hep.26130]

Lao P, Yin P, Hua R, Tan Y, Li Z, Qu Y, Qin Z, Xie Y, Xie X, Wang X, Chen Z, Zhou L, Wang X, Li Y, Chen H, Gao L, Xu X, Wu T, Wang H, Niu J, Xu G. A Large-scale, multicenter serum metabolite biomarker identification study for the early detection of hepatocellular carcinoma. Hepatology 2018; 67: 662-675 [PMID: 28963740 DOI: 10.1002/hep.29561]

Huang Q, Tan Y, Yin P, Ye G, Gao P, Lu X, Wang H, Xu G. Metabolic characterization of hepatocellular carcinoma using nontargeted tissue metabolomics. Cancer Res 2013; 73: 4909-5002 [PMID: 23824744 DOI: 10.1158/0008-5472.CAN-13-0308]

Zhou P, Li J, Shao L, Lv G, Zhao L, Huang H, Zhang A, Pan X, Liu W, Xie Q, Chen D, Guo Y, Hao S, Xu W, Li L. Dynamic Patterns of serum metabolites in fullminant hepatic failure pigs. Metabolomics 2012; 8: 869-879 [DOI: 10.1007/s11306-011-0381-5]

Zhou P, Zhou N, Shao L, Li J, Liu S, Meng X, Duan J, Xiong X, Huang X, Chen Y, Fan X, Zheng Y, Ma S, Li C, Wu A. Diagnosis of Clostridium difficile infection using an UPLC-MS based metabolomics method. Metabolomics 2018; 14: 102 [PMID: 30830376 DOI: 10.1007/s11306-018-1397-x]

Zhou J, Sun HC, Wang Z, Cong WM, Wang JH, Zeng MS, Yang JM, Bie P, Liu LX, Wen TF, Han GH, Wang MQ, Liu RB, Lu LG, Ren ZG, Chen MS, Zeng ZC, Liang P, Liang CH, Chen M, Yan FH, Wang WP, Ji Y, Cheng WW, Dai CL, Jia WD, Li YM, Li YX, Liang J, Liu TS, Lv GY, Mao YL, Ren WX, Shi HC, Wang WT, Wang XY, Xing BC, Xu JM, Yang JY, Yang YF, Ye SL, Yin ZY, Zhang BH, Zhang SJ, Zhou WP, Zha JY, Liu R, Shi YH, Xiao YS, Dai Z, Teng GJ, Cai JQ, Wang WL, Dong JH, Li Q, Shen F, Qin SK, Fan J. Guidelines for Diagnosis and Treatment of Primary Liver Cancer in China (2017 Edition). Liver Cancer 2018; 7: 235-260 [PMID: 30319983 DOI: 10.1111/livc.13083]

Qin N, Yang F, Li A, Prihi E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich SD, Zheng S, Li L. Alterations of the human gut microbiome in liver cirrhosis. Nature 2014; 513: 59-64 [PMID: 25079328 DOI: 10.1038/nature13568]

Chen E, Lu J, Chen D, Zhu D, Wang Y, Zhang Z, Zhou N, Wang B, Li J, Li L. Dynamic changes of plasma metabolites in pigs with GalN-induced acute liver failure using GC-MS and UPLC-MS. Biomed Pharmacother 2017: 93: 480-489 [PMID: 28686767 DOI: 10.1016/j.biopha.2017.06.049]

Chen R, Zhu S, Fan XG, Wang H, Lotze MT, Zhe HJ 3rd, Billiar TR, Kang R, Tang D. High mobility group protein B1 controls liver cancer initiation through yes-associated protein-dependent aerobic glycolysis. Hepatology 2018; 67: 1823-1841 [PMID: 29149457 DOI: 10.1002/hep.29663]

Fittan AI, Cabrera R. Disease monitoring of hepatocellular carcinoma through metabolomics. World J Hepatol 2017; 9: 1-17 [PMID: 28105253 DOI: 10.4245/wjg.v9.i1.1]

Wise DR, Thompson CB. Glutamine addiction: a new therapeutic target in cancer. Trends Biochem Sci 2010; 35: 423-437 [PMID: 20570523 DOI: 10.1016/j.tibs.2010.05.003]

Tremosini S, Forner A, Boix L, Vilana R, Reich M, Rimola J, Rodríguez-Lope C, Ayuso C, Solé M, Bruix J. Prospective validation of an immunohistochemical panel (glycpanic 3, heat shock protein 70 and glutamine synthetase) in liver biopsies for diagnosis of very early hepatocellular carcinoma. Gut 2012; 61: 1481-1487 [PMID: 22287594 DOI: 10.1136/gutjnl-2011-301862]

Howard SC, Jones DP, Pui CH. The tumor lysis syndrome. N Engl J Med 2011; 364: 1844-1854 [PMID: 21561350 DOI: 10.1056/NEJMoa0904569]

McPhail MJW, Shaweross DL, Lewis MR, Coltart I, Want EI, Antoniades CG, Veselkov K. Triatomicfylou E, Patel V, Pop G, Gomez-Romo MR, Kyräskides M, Zia R, Aebles RD, Crossley MME, Jassem W, O'Grady P, Heaton N, Auinger G, Bernal W, Quaglia A, Coen M, Nicholson JK, Wendorf JA, Holmes E, Taylor-Blaxton SD. Multivariate metabolic profiling of plasma predicts survival in patients with decompensated cirrhosis. J Hepatol 2016; 64: 1058-1067 [PMID: 26795831 DOI: 10.1016/j.jhep.2016.01.003]
Zhou PC et al. Metabolomics model for screening HCC.

metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort. J Proteome Res 2012; 11: 5914-5923 [PMID: 23078175 DOI: 10.1021/pr300673x]
