Frontiers in Bioscience Landmark

*Front. Biosci. (Landmark Ed)* **2024**; 29(1): 3
https://doi.org/10.31083/j.fbl2901003

*Original Research*

# Protein Signatures for Distinguishing Colorectal Cancer Liver Metastases from Primary Liver Cancer Using Tissue Slide Proteomics

Xiaoman Zhou[1], Xiuyuan Wang[1], Ruizhen Bai[2], Hanjie Li[1], Dong Hua[3], Xiao-Dong Gao[4], Ganglong Yang[1,4,*], Quan Liu[5,*]

[1]The Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, School of Biotechnology, Jiangnan University, 214122 Wuxi, Jiangsu, China
[2]Department of Pathology, Affiliated Hospital of Jiangnan University, 214122 Wuxi, Jiangsu, China
[3]Department of Oncology, Wuxi People's Hospital Affiliated to Nanjing Medical University, 214122 Wuxi, Jiangsu, China
[4]State Key Laboratory of Biochemical Engineering, Institute of Process Engineering, Chinese Academy of Sciences, 100190 Beijing, China
[5]Department of Medical Oncology, Affiliated Hospital of Jiangnan University, 214122 Wuxi Jiangsu, China
*Correspondence: glyang@ipe.ac.cn (Ganglong Yang); quanliu@jiangnan.edu.cn (Quan Liu)
Academic Editor: Zhaoguo Liu
Submitted: 6 June 2023 Revised: 7 September 2023 Accepted: 15 September 2023 Published: 9 January 2024

**Abstract**

**Background**: Colorectal cancer liver metastasis (CRLM) and hepatocellular carcinoma (HCC) are both high incidence tumors in China. In certain poorly differentiated cases they can exhibit comparable imaging and pathological characteristics, which impedes accurate clinical diagnosis. The use of protein-based techniques with tissue slides offers a more precise means to assess pathological changes and has the potential to assist with tumor diagnosis. **Methods**: A simple *in situ* protein digestion protocol was established for protein fingerprint analysis of paraffin-embedded tissue slide samples. Additionally, machine learning techniques were employed to construct predictive models for CRLM and HCC. The accuracy of these models was validated using tissue slides and a clinical database. **Results**: Analysis of differential protein expression between CRLM and HCC groups reliably identified 977 proteins. Among these, 53 were highly abundant in CRLM samples and 57 were highly abundant in HCC samples. A prediction model based on the expression of six proteins (CD9, GSTA1, KRT20, COL1A2, AKR1C3, and HIST2H2BD) had an area under curve (AUC) of 0.9667. This was further refined to three proteins (CD9, ALDH1A1, and GSTA1) with an AUC of 0.9333. **Conclusions**: Tissue slide proteomics can facilitate accurate differentiation between CRLM and HCC. This methodology holds great promise for improving clinical tumor diagnosis and for identifying novel markers for challenging pathological specimens.

**Keywords:** hepatocellular carcinoma; colorectal cancer liver metastasis; proteomics; FFPE tissue slide

## 1. Introduction

The liver is the primary metabolic organ in the body and also a frequent site of tumor occurrence. In addition to primary liver cancer (hepatocellular carcinoma, HCC), liver tumors often originate from other tissues and are therefore known as secondary liver cancers. The liver's unique physiology, characterized by a high blood supply, makes it an ideal site for tumor colonization [1–3]. Colorectal cancer (CRC) is the most common of the different primary tumor types that metastasize to the liver. More than half of all high-grade CRC cases exhibit liver metastases, even after removal of the primary tumor [4,5]. In certain cases, the histological characteristics of poorly differentiated cancer cells pose a challenge in terms of distinguishing between primary and metastatic tumors [6,7]. While medical history can help to exclude certain tumor types, there are instances where the primary tumor is not obvious, thereby making the diagnosis more challenging. Distinguishing between liver metastases from CRC and HCC is critical for determining the appropriate treatment option and thus achieving better patient survival [8].

The early detection and diagnosis of CRC liver metastasis (CRLM) is important for improving patient outcome and increasing the likelihood of successful treatment [2,9,10]. Certain blood markers such as myeloperoxidase (MPO), carcinoembryonic antigen (CEA), and carbohydrate antigen (CA) 19-9 can potentially indicate the existence of CRLM [5]. However, these markers are not exclusive to CRLM and are not sufficiently accurate [11,12]. While CEA is commonly employed to predict CRC, its sensitivity as a biomarker for CRLM is unsatisfactory [11]. The insufficient specificity and sensitivity of the current biomarkers necessitates the discovery of novel biomarkers [13]. Although transcriptome analysis of tumor tissue is considered to be the gold standard for distinguishing between CRLM and HCC and is recommended for clinical use, rapid pathological testing based on histology has the advantage of being conducted in-hospital. This helps with the prompt selection of a treatment plan for postoperative patients. However, in some poorly differentiated cases, there is still a lack of accurate immunohistochemical markers that can accurately differentiate HCC from CRLM in routine pathological tests.

Due to their extended storage, formalin-fixed paraffin-embedded (FFPE) tissue slices often closely resemble the original clinical diagnostic sample [14]. Hence, they are very appropriate for the discovery of novel biomarkers. The field of FFPE proteomics involves the study of proteins extracted from FFPE tissue samples. This research area has gained considerable importance in recent years [15–17]. Improved techniques for sample preparation, as well as advances in mass spectrometry-based proteomics, have significantly increased the accuracy and sensitivity of protein analysis from FFPE samples. In addition, accessible computational platforms for big data analysis and artificial intelligence (AI) technologies can improve proteomic profiling and biomarker screening, thereby advancing the translational application of proteomics data in the field of precision medicine [13]. In previous studies, we developed a rapid proteomics and glycomics method to analyze tissue samples [14,15]. We now apply this sample preparation method for rapid proteomics using FFPE tissue slides in order to build a model for tumor identification using AI.

In this study, we used FFPE slides from CRC, CRLM, and HCC patients to carry out rapid proteomic analysis. Subsequently, we conducted differential protein intensity and functional analyses based on the proteomics results. A machine learning model was then constructed to simplify the diagnostic model, and immunohistochemistry subsequently used to validate the CRLM and HCC differential detection markers.

## 2. Materials and Methods

### *2.1. Tissue Samples*

FFPE tissue specimens from 54 patients who underwent surgery at the Affiliated Hospital of Jiangnan University were selected for analysis. This study complied with the basic principles of medical ethics in the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of the Affiliated Hospital of Jiangnan University (NO. LS2021-048). Patients or their relatives were fully informed and signed an informed consent form. Patient IDs and other personal information were coded in order to protect patient privacy. The clinicopathological information associated with these specimens is shown in **Supplementary Table 1**. FFPE tissue sections (4 µm) from each patient were stained with hematoxylin/eosin and evaluated by a pathologist to determine the tumor type and to assess tumor grade. The samples used in the current study consisted of 22 CRC samples, 19 HCC samples, and 13 CRLM samples.

### *2.2 Tissue Slide Treatment*

Tissue slides were first baked at 72 °C for 30 min, followed by deparaffinization in xylene (#10023418, Sinopharm chemical reagent, Shanghai, China), rehydration through a graded ethanol (#100091192, Sinopharm chemical reagenta) series, and boiling in a citrate antigen retrieval solution (#P0086, Beyotime, Shanghai, China) for protein recovery. They were then washed twice in dd$H_2O$ and dried *in vacuo*.

### *2.3 Protein Digestion on Slides*

The on-slide tissue digestion protocol was from a previously published protocol [18]. Tissue slides were kept in a wet box for the subsequent step. First, tissue was reduced with 50 µL of 10 mM dithiothreitol (DTT) solution (#D0632, Sigma Aldrich; St. Louis, MO, USA) per square centimeter and incubated at 60 °C for 30 minutes. Subsequently, 50 µL of 20 mmol/L iodoacetamide (IAM, #I6125, Sigma Aldrich) was added for alkylation and the slides kept in the dark at room temperature for 45 min. Next, 120 µL of digestion buffer (40 mmol/L $NH_4HCO_3$, 10 µg trypsin (#TRY001C, sequencing grade, Shengxia; Beijing, China)) was added to the tissue and incubated overnight at 37 °C. Finally, 60 µL of 40 mmol/L $NH_4HCO_3$ was added for 3 min, and pipette tips were used to collect the peptide fragments into a 1.5 mL tube. The process was repeated three times to obtain all of the peptide solution. C18 pipet tips (ZipTip, Millipore Corp; Billerica, MA, USA) were employed to desalt the peptides.

### *2.4 Tandem Mass Tag (TMT) Labeling of Peptides*

Peptides from the tissue slide were labeled with TMT reagent according to a previously described protocol [19]. The required volume to obtain 1.5 µg of peptides per sample was dried *in vacuo* and resuspended in 50 µL of 100 mmol/L 4-hydroxyethylpiperazine ethanesulfonic acid (HEPES) buffer (pH 8.5). TMT10plex labeling reagents (#90309, Thermo Fisher Scientific; MA, USA) were dissolved in anhydrous acetonitrile and added to the peptide sample in a 1:10 peptide/TMT ratio [20]. Samples were incubated for 2 h (22 °C, 750 rpm) and unreacted TMT reagent was quenched by incubation with 5 µL of 5% hydroxylamine (#467804, Sigma Aldrich) for 35 min. Samples belonging to the same batch (**Supplementary Table 2**) were combined and dried *in vacuo*.

### *2.5 LC-MS Analysis*

Lyophilized peptides were reconstituted in an aqueous solution comprised of 2% acetonitrile and 0.1% formic acid (#1.00029, #5.33002, Sigma Aldrich) and then analyzed by Orbitrap Fusion Lumos mass spectrometry (Thermo Scientific, San Jose, CA, USA). Each test involved the injection of 2 µg of mixed TMT samples, with two replicates performed for each proteomics experiment. Samples were then separated using an EASY-nLC 1200 system with a C18 column (75 µm × 25 cm) (Thermo Scientific). Other mass spectrometry parameters were the same as those of established methods used in our laboratory [21,22].

*2.6 Protein Identification and Data Analysis*

The tandem mass spectra of peptide samples were searched against the Uniprot Homo sapiens protein sequence database (version: May 2022) using MaxQuant software (Max-Planck-Institute of Biochemistry, Martinsried, Germany, version 2.3.1.0) [23]. The cleavage enzyme was set to trypsin/P, allowing for up to two misses, and the mass tolerances for the precursor ions and MS2 fragment ions were set to 4.5 ppm and 20 ppm, respectively. Fixed modification was set to carbamidomethylation of cysteine, and variable modifications were set to the oxidation of methionine and the methylation of lysine [24]. The search was set to reporter ion MS2 using TMT10plex. All twelve raw files from the six TMT batches were analyzed in parallel.

The results were visualized using R (version 4.2.3, https://www.r-project.org/) and packages that included ggpubr (version 0.6.0), plotROC (version 2.3.0) [25], ggplot2 (version 3.4.1), and ComplexHeatmap (version 2.15.2) [26]. Gene Ontology (GO) enrichment was performed using the R package clusterProfiler (version 4.6.2) [27]. Seurat (version 4.3.0) was used to process the spatial transcriptome and scRNA data [28]. Samples were clustered by unsupervised clustering based on the abundance of proteins, using the Ward.D method in the package cluster of R. Student's *t*-test was used to evaluate the statistical significance of differences between groups (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; "ns", not significant).

*2.7 LASSO Regression Analysis*

Feature selection was performed by applying least absolute shrinkage and selection operator (LASSO) regression to the significant proteins identified by the comparative proteomic analysis. In all, 56 samples were randomly divided into 70% for model training and 30% for model evaluation. A logistic regression model was fitted with the clinical diagnosis (CRLM = 1, HCC = 0). Receiver operating characteristic (ROC) analysis was then applied to the predicted probability and actual diagnosis, respectively. All analyses were performed using RStudio (version 2023.03.0, Posit, Boston, MA, USA) with R (4.2.3). LASSO regression was performed using the glmnet package (4.1) [29], and ROC analysis was performed using the plotROC package (2.3.0) [25].

*2.8 Immunohistochemistry*

The immunohistochemistry protocol used here was based on a previous publication [30]. FFPE slides were prebaked at 72 °C for 30 min before undergoing deparaffinization in xylene and rehydration in a graded series of ethanol. The slides were then boiled in EDTA 9.0 solution to obtain antigen recovery. Endogenous peroxidase was blocked in 3% hydrogen peroxide for 15 min, followed by a wash in PBS. The slides were then incubated overnight at 4 °C with primary antibodies against CD9 (1:500, #ab263019, Abcam, Shanghai, China), ALDH1A1 (1:200, #ab52492, Abcam), or GSTA1 (1:400, #14475, Proteintech, Wuhan, China). They were then washed in PBS and incubated with secondary antibodies at room temperature for 20 min. DAB chromogenic solution (#P0203, Beyotime, Shanghai, China) was used to visualize the staining according to the manufacturer's instructions. Cell nuclei were counterstained with hematoxylin, followed by dehydration, hyalinization in xylene, and sealing for microscope viewing.

## 3. Results

*3.1 Sample Processing and Data Processing for FFPE Proteomics*

FFPE tissue slides from 19 HCC patients, 13 CRLM patients, and 22 CRC patients were used in the proteomics experiments. The sample preparation workflow is outlined in Fig. 1A. Tissue samples were carefully examined by pathology experts to ensure accurate tissue classification and quality control (Fig. 1B). Following digestion, the purified peptides were collected, labeled with tandem mass tags (TMTtags), and then mixed based on a random design table (**Supplementary Table 2**). Following protein identification and quantification using MaxQuant software, data was combined using the reference channel (TMT_10) and data normalization was then performed (**Supplementary Fig. 1A,B**). A total of 2628 proteins were detected across all sample replicates. After eliminating proteins with identification scores <10 and detection rates <70%, 977 reliable proteins were retained for further analysis (**Supplementary Fig. 2A,B**). The abundance ranking of these filtered proteins was compared to all detected proteins, with most of the selected proteins having high abundance (**Supplementary Fig. 2B**).

*3.2 Overall Analysis of Proteins Detected by Tissue Slide Proteomics*

To gain insight into the proteins detected in the tissue slide samples, a comprehensive statistical analysis was conducted for all reliable proteins (Fig. 1C). The overall protein intensity distribution was similar across the three sample groups (Fig. 1C), but several significant differences were apparent (**Supplementary Fig. 3**). A statistical analysis was performed on the detected proteins based on their subcellular localization. Cytoplasmic and nuclear proteins comprised most of the total protein (**Supplementary Fig. 4A,B**). A protein functional annotation analysis was performed to analyze the functions of all reliable proteins detected in the three types of tumor tissue. These proteins were primarily associated with cell focal adhesion, the transport of vesicles, the junction between cells and substrate, metabolism pathways such as cadherin binding, structural constituents of the ribosome, and actin filament binding (**Supplementary Fig. 4C**). In cluster analysis, the CRLM and HCC groups were found to be more similar, whereas the CRC group was readily distinguishable from both the CRLM and HCC groups (Fig. 1D).

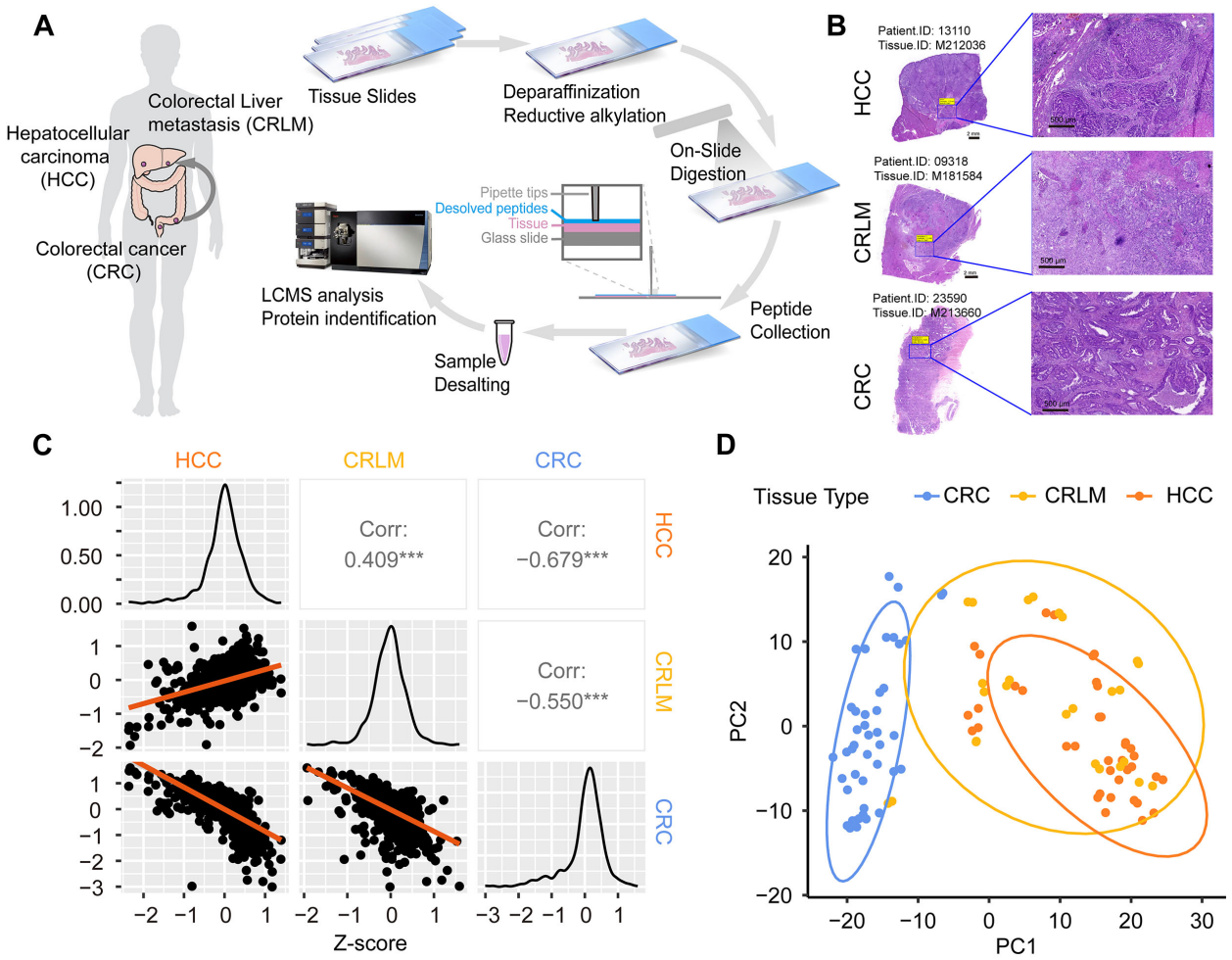


**Fig. 1. Sample information and sample processing workflow.** (A) Workflow for the tissue-slide-based proteomics analysis. (B) Hematoxylin–eosin (HE) staining of tissue slides for pathological diagnosis. Scale bar: 2 mm (left), 500 µm (right). (C) Correlation analysis for the three sample groups. Data are presented as a matrix, including correlation dot plot (lower), density distribution (diagonal), and correlation coefficient with significance markers (upper). The correlation coefficient was calculated by Pearson correlation analysis. *** denotes $p < 0.001$. (D) Principal component analysis (PCA) for proteins detected in colorectal cancer (CRC), CRC liver metastasis (CRLM), and hepatocellular carcinoma (HCC) tissue samples.

A trend cluster analysis was performed to identify protein differences between the three tumor types (Fig. 2). The proteins were classified into five clusters based on their intensity. Cluster 3 was comprised of proteins that were high in HCC but low in CRLM and CRC, with functions primarily focused on protein folding and degradation. Cluster 1 included proteins that were high in CRLM and HCC but low in CRC, with functions concentrated mainly on small molecule and energy metabolism pathways. Cluster 5 consisted of proteins that were highest in CRLM, but relatively low in HCC and CRC, and were primarily associated with cytoskeletal cell formation and migration. Clusters 2 and 4 proteins were higher in CRC, but relatively low in HCC and CRLM, with functions mainly concentrated on generating cytoplasmic proteins and mediating cell adhesion.

### *3.3 Comparative Proteomic Analysis of CRLM and HCC*

Differential protein intensity analysis was performed between the CRLM and HCC sample groups in order to identify protein signatures for different tumor types. The results are plotted as a volcano plot in Fig. 3A. A total of 110 differentially expressed proteins were detected, with 53 that were significantly up-regulated in CRLM and 57 considerably up-regulated in HCC (Fig. 3A). Moreover, results from the functional annotation analysis revealed that CRLM-enriched proteins were mainly related to collagen and participated in epithelial tissue formation, tissue repair, cell migration, and physiological activities such as cell adhesion and extracellular matrix remodeling (Fig. 3B,D). Conversely, proteins that were significantly increased in HCC were primarily associated with cell energy and substance metabolism, and to the synthesis and metabolism of organic compounds (Fig. 3C,D).

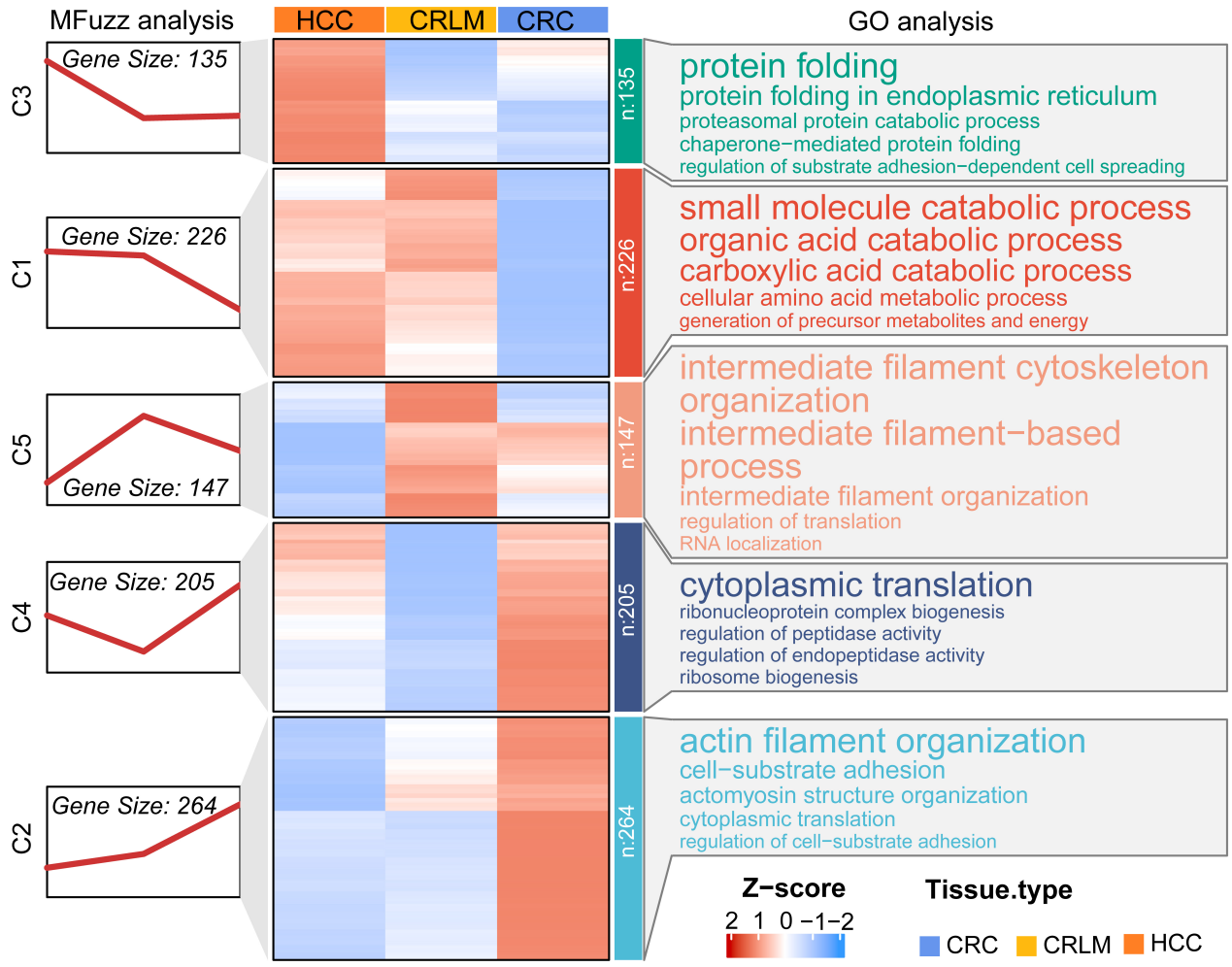


**Fig. 2. Cluster analysis for protein intensity in three tissue slide groups.** The mean of the z-score normalized protein intensity is shown as a heatmap (middle). Column: sample group. Row: protein name divided into five groups by MFuzz clustering and arranged by cluster. Five clusters were grouped by trend clustering (left). Molecular function annotation of the five clusters (right), with the font size related to the enrichment coefficient. GO, Gene Ontology.

To identify proteins that can distinguish CRLM from HCC, the screening criteria was increased to a fold change >2.83 ($\log_2$ (foldchange) >1.5) and an adjusted *p*-value of <0.005. This identified 16 proteins, with 10 proteins significantly increased in CRLM, and 6 significantly increased in HCC (Fig. 4A). The ROC results showed the proteins that were increased in CRLM served as positive markers, with area under curve (AUC) values >0.75. Conversely, by using proteins that were increased in HCC as negative indicators for the diagnosis of CRLM, the AUC values were >0.75 (Fig. 4B).

### *3.4 Development of a Model for Discriminating Between CRLM and HCC*

Detection using multiple proteins was then performed to increase the accuracy when distinguishing between CRLM and HCC (Fig. 5A). Protein intensity data from the selected samples was randomly divided into training and validation groups at a ratio of 7:3, respectively (Fig. 5B). A binomial classification LASSO regression model was established for detection. When the lambda value was smallest (Lambda.min), this indicated the most accurate model, and six variables were selected for prediction. Ranked by weight, these were CD9, glutathione S-transferase A1 (GSTA1), keratin type I cytoskeletal 20 (KRT20), collagen alpha-2(I) chain (COL1A2), aldo-keto reductase family 1 member C3 (AKR1C3), and putative histone H2B type 2-D (HIST2H2BD) (Fig. 5C,D). When the lambda value was optimal (Lambda.1se), this indicated the simplest model, and three variables (CD9, GSTA1, and aldehyde dehydrogenase 1A1 (ALDH1A1)) were selected for prediction (Fig. 5C,D). Using the accurate model based on six proteins, the AUC value of the ROC curve for predicting sample classification was 0.9667, while the simplified model based on three proteins had an AUC value of 0.9333 (Fig. 5E). The scores obtained using the accurate model to diagnose CRLM and HCC were very close to the actual pathological diagnosis. Most of the correct results were still obtained using the simplified model (Fig. 5F).

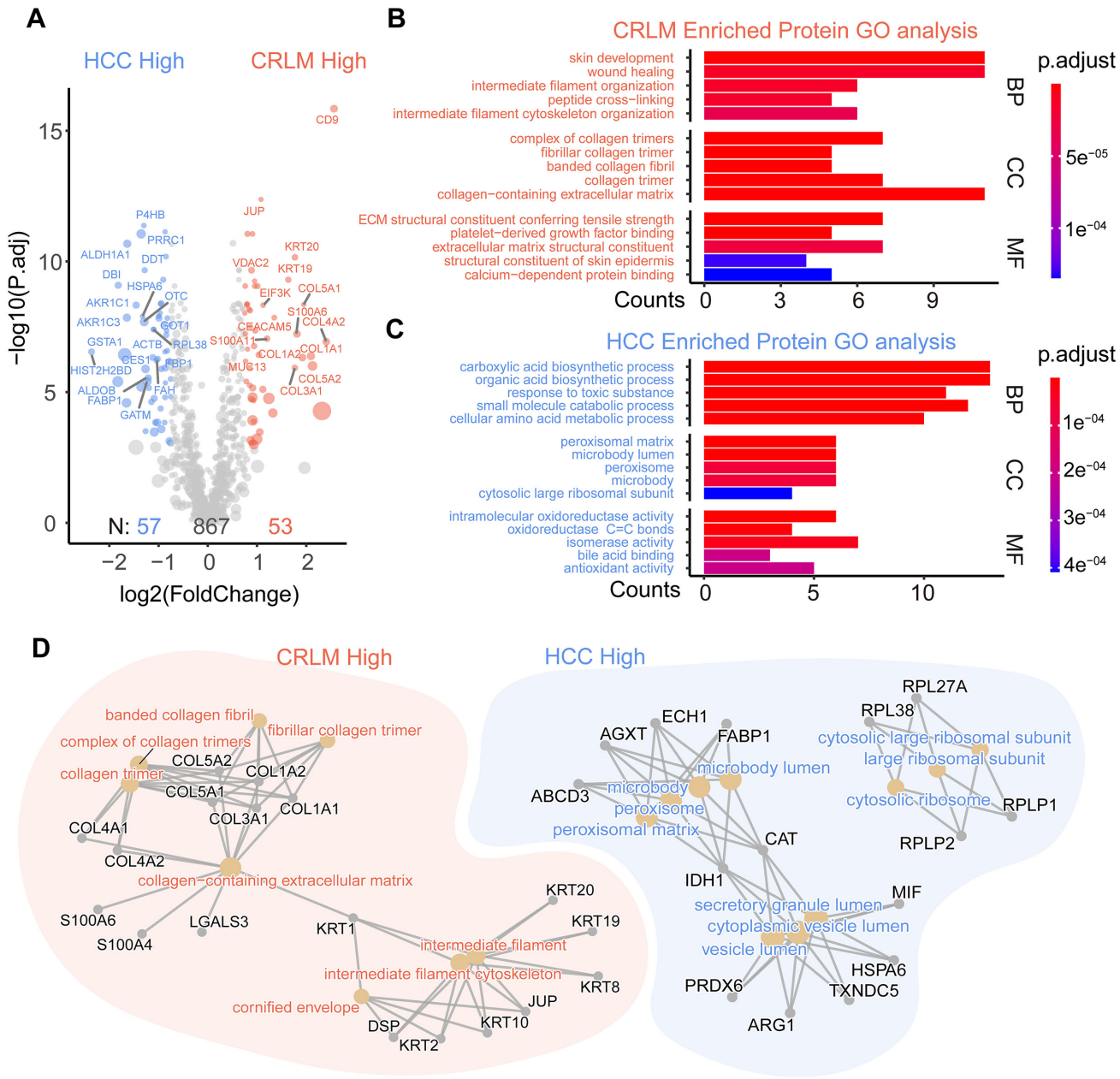


**Fig. 3. Differential intensity analysis of proteins between the CRLM and HCC groups.** (A) Volcano plot of proteins, with fold-change >2 and *p*.adjust < 0.001 set to identify significant proteins between the two groups. GO analysis of significantly increased proteins in the CRLM (B) and HCC (C) groups. (D) Cellular component enrichment network for the CRLM and HCC groups.

### *3.5 Tissue Validation of the Proteins in the Prediction Model*

We have confirmed the identification results of the marker proteins in the mass spectrometry search results. The raw mass spectrometry files and search results have been uploaded to ProteomeXchange database (PXD042636), with representative spectra shown in **Supplementary Fig. 5**. To further validate the actual status of the signature proteins in clinical tumor detection, we conducted verification using publicly available databases and pathological tissue samples. We conducted tissue immunostaining to validate the CD9, ALDH1A1 and GSTA1 biomarkers (Fig. 6). CD9 showed strong positive staining in the cancerous region of CRLM, but was negative in HCC. ALDH1A1 and GSTA1 showed strong positive staining in HCC, but did not stain in CRLM (Fig. 6A). We also examined the CD9, ALDH1A1, and GSTA1 mRNA level in the single-cell RNA and spatial transcriptome dataset. The results of this analysis were consistent with those of immunohistochemistry (Fig. 6B, **Supplementary Figs. 6,7**). Taken together, these findings demonstrate the CD9, ALDH1A1 and GSTA1 biomarkers can reliably distinguish HCC from CRLM.

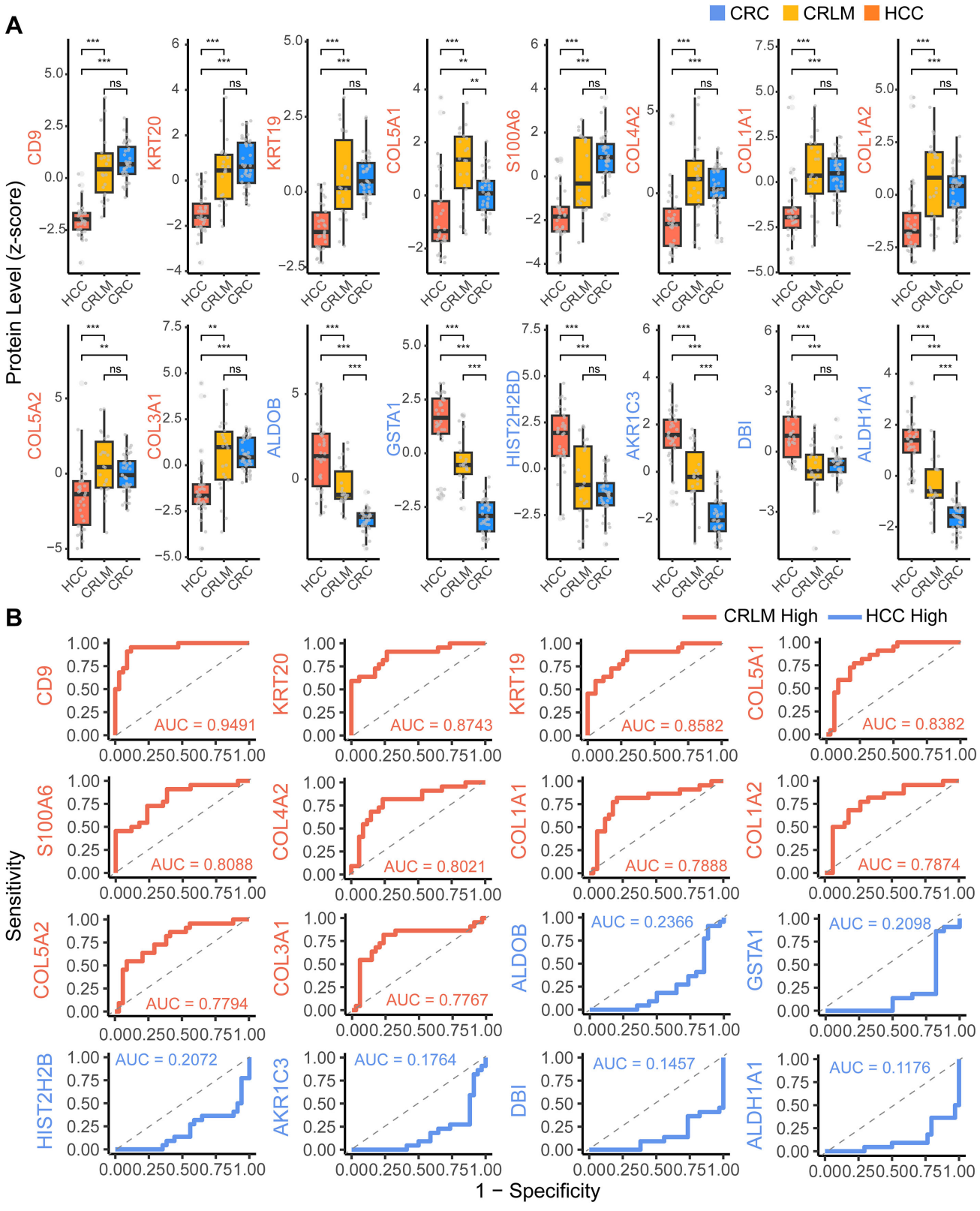


**Fig. 4. Significance of differentially expressed proteins.** (A) Box plot with data points and significance analysis of differentially expressed proteins in three tissue groups. Student's *t*-test was used to evaluate the statistical significance. "ns" denotes not significant, ** denotes $p < 0.01$, *** denotes $p < 0.001$, and "ns" denotes not significant. (B) Receiver operating characteristic (ROC) analysis of differentially expressed proteins as diagnostic markers for CRLM. The area under the curve (AUC) was used to evaluate the specificity and sensitivity of these proteins. Proteins shown in red were expressed at high levels in CRLM, while those shown in blue were expressed at high levels in HCC.

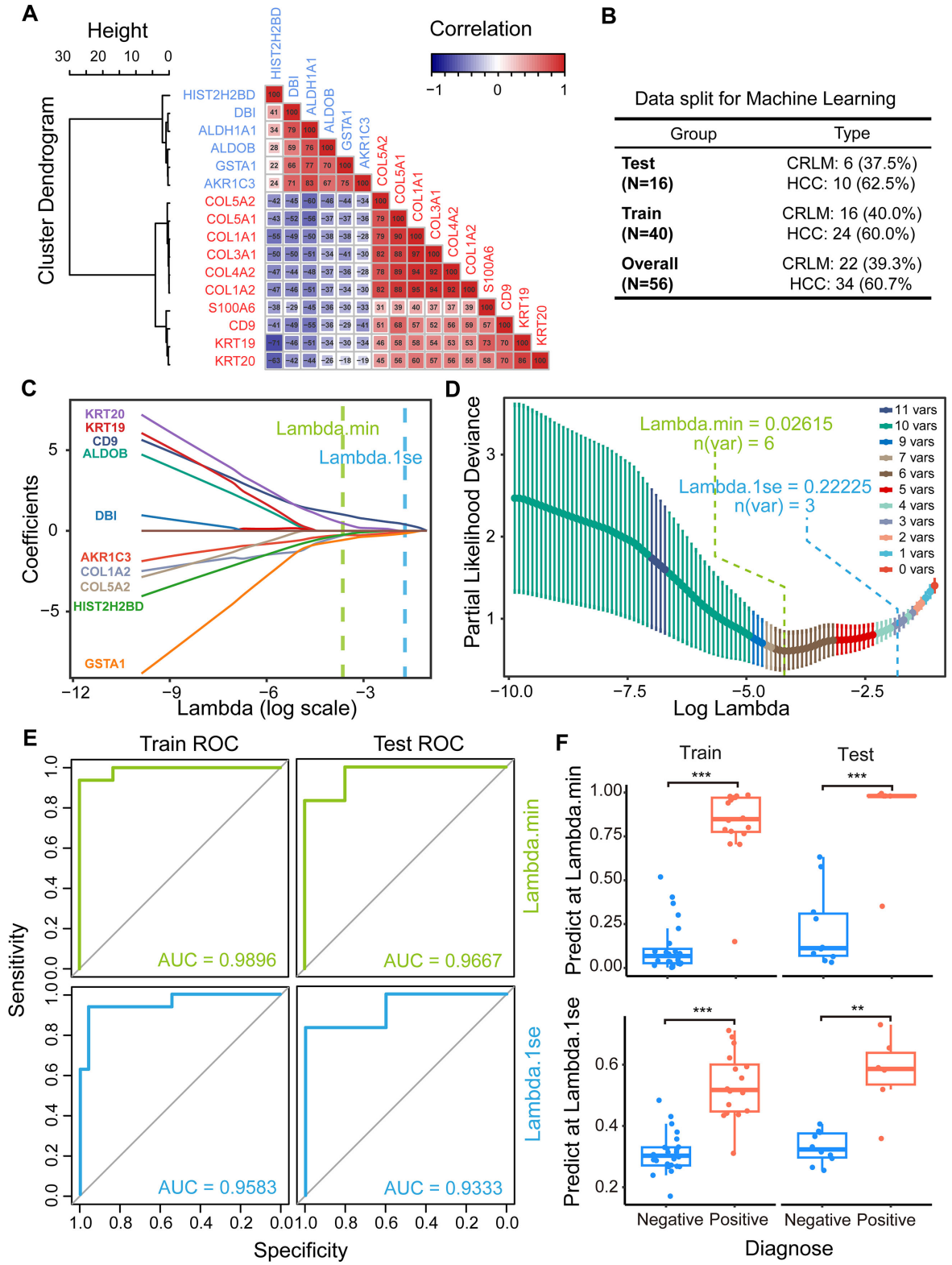


**Fig. 5. Machine learning (ML)-based model for CRLM diagnosis.** (A) Correlation matrix of differentially expressed proteins between CRLM and HCC. (B) Table of sample separation for ML. The training group (70%) and test group (30%) were used to build and evaluate the prediction model, respectively. (C) Lasso coefficient profiles of all the clinical features. The lasso regression model and cross-validation method were used to choose proteins for diagnosis. (D) Identification of the optimal penalization coefficient $\lambda$ in the Lasso mode. Dotted vertical lines were drawn at the optimal values using the minimum criteria (min) and the 1 standard error (1se) of the minimum criteria. (E) ROC analysis of the CRLM diagnosis model. (F) Box plot of prediction value and actual diagnoses. Student's *t*-test was used to evaluate the statistical significance. ** denotes $p < 0.01$, *** denotes $p < 0.001$.

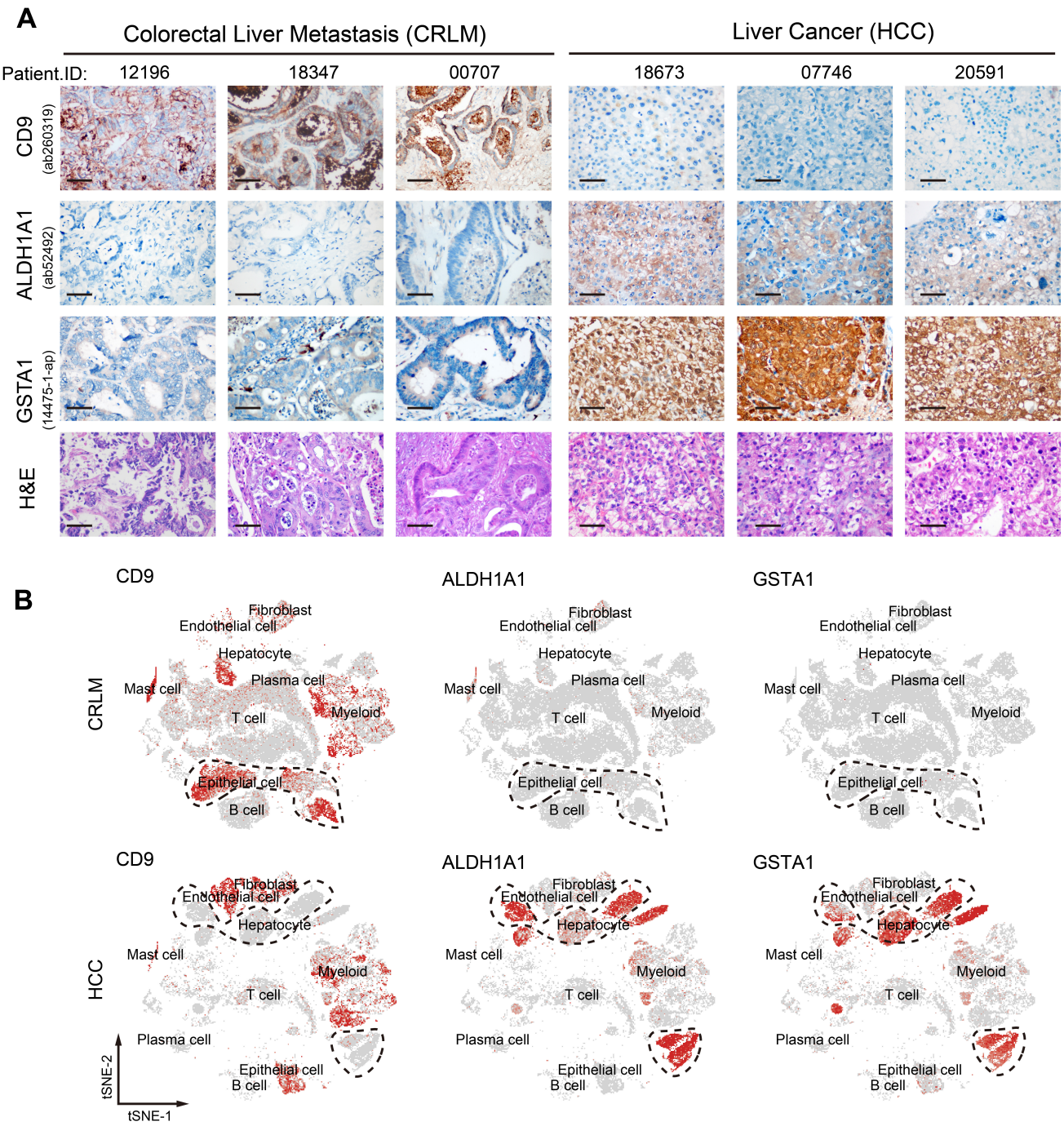


**Fig. 6. Validation of the proteins obtained in the prediction model.** (A) Immunohistochemical staining of CD9, ALDH1A1 and GSTA1 in tissue slides from CRLM and HCC patients. Scale bar: 50 μm. (B) Single-cell RNA (scRNA) analysis of CD9, ALDH1A1 and GSTA1 in tumor cells from CRLM and HCC (dashed circle). The data are from GSE178318 and GSE149614. The red dot indicates the expression intensity of the corresponding gene in the cell.

## 4. Discussion

CRC has a high propensity for liver metastasis and is often difficult to differentiate from HCC, especially in some poorly differentiated cases. Furthermore, both HCC and CRLM have a high incidence rate in China, with cases of dual tumors coexisting. Therefore, it is important to accurately differentiate between these two cancer types for optimal patient diagnosis and treatment [31,32]. Patients with CRLM are usually in advanced stages, and the majority are not suitable candidates for surgery. This presents a major challenge in the acquisition of fresh tissue samples for routine proteomic analysis. FFPE samples are an ideal source for CRLM research since they are commonly used for tumor detection in pathology departments and have stable properties with long preservation times. The current study used FFPE samples for proteomic analysis, allowing

samples that were already present in the pathology tissue bank to be used and thus saving time and resources required for sample collection [33]. Furthermore, characteristic proteins detected in samples obtained from tissue slices are more suitable in subsequent diagnosis and testing of slice samples.

After years of development, the sample preparation techniques used in proteomics and the depth of detection have undergone significant improvement. The identification of more proteins, particularly those present in low concentrations, is essential for elucidating biological processes and tumor metabolism [33–36]. Following tissue fixation with formaldehyde, biopsies are usually archived in paraffin blocks. Currently, two main factors affect the use of FFPE slides for large-scale proteomic studies. The first is limited protein extraction due to poor solubility, and the second is uncertain protein identification due to possible unknown peptide modifications. Formaldehyde alters the chemical composition of proteins by causing intra- and inter-protein cross-links, as well as links between proteins and nucleic acids. Such alterations make protein analysis by mass spectrometry considerably more challenging [14,37]. Those modification also have negative effects on protein digestion by trypsin, and on peptide detection during database screening [38]. However, the formation of some formaldehyde adducts and cross-links was found to be reversible at elevated temperatures or pressure. Heat treatment during sample preparation helps to improve detection [39]. For clinical detection, and especially pathological detection, a relatively high abundance of protein makes it easier to successfully perform tissue immunohistochemistry using specific antibodies. Taking advantage of tissue slide proteomics, we used on-slide protein sample reduction, alkylation, and digestion during sample preparation for mass spectrometry that does not rely on detergents. This significantly reduces the overall sample preparation time and minimizes sample loss. Using this protocol, we reliably identified a total of 977 proteins. Following an abundance ranking analysis of these filtered proteins compared to identified proteins, we concluded the identified proteins mainly had high abundance. This could be because no strong detergents or complex sample preparation methods were used during our mass spectrometry analysis [40,41]. Current mainstream FFPE proteomics methods typically use scraping or laser-cutting to transfer tissue from the glass slides and into centrifuge tubes, followed by sample preparation methods involving detergent treatment and enzymatic digestion in solution. Comparison of these methods showed that a similar number of proteins was detected [42]. The use of strong detergents can more effectively denature proteins and disrupt protein tertiary structures, which helps to improve the efficiency of subsequent protein digestion [33]. However, complex sample preparation processes require high operator proficiency and are not conducive to automation. Residual detergents can also affect protein identification by instruments. Our detected proteins overlapped with high-abundance proteins reported by other related research [43], thus demonstrating the reliability of our method for identifying major proteins. Moreover, our approach has the advantages of being compatible with automated sample preparation processes based on liquid handling stations, as well as sample detection in subsequent spatial tissue proteomics based on mass spectrometry [44].

To differentiate between FFPE CRLM and HCC samples using proteomics analysis, we conducted a differential protein-intensity analysis of these two cancer types. The results showed that enriched proteins in CRLM were mainly related to the extracellular matrix and to cell movement, whereas those in HCC were primarily associated with energy and material metabolism (Figs. 2,3). The results for CRLM were consistent with the features of tumor metastasis, and the identified collagen-related proteins (COL1A1, COL1A2, COL4A1, COL4A2, COL5A1, COL5A2) were previously confirmed as marker proteins of CRLM [43]. Following metastasis of CRC to the liver, the tumor microenvironment undergoes reshaping. This affects the composition of both the extracellular matrix and immune cells, while reducing the efficacy of targeted therapies such as CAR-T. Moreover, it has been reported previously that MUC13 [45], carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5) [46,47], and $Ca^{2+}$-binding proteins (S100A6 and S100A11) [48] are highly expressed in CRLM and can thus be used as diagnostic markers. Several highly expressed proteins in HCC, such as heat shock 70 kDa protein 6 (HSPA6) [49], protein disulfide-isomerase (P4HB), and catalase (CAT) [50] have also been used for the detection of this cancer type. The above results indicate that our FFPE rapid proteomic profiling method can be used for the differential diagnoses of CRLM and HCC.

In the present study we also employed a simple LASSO regression model for binary classification prediction using 16 selected detection markers. This approach helped to reduce the number of detection markers, while still maintaining accuracy and sensitivity. Finally, we built a simple prediction model using only three variables (CD9, GSTA1, and ALDH1A1). These three biomarkers are typical proteins found in the primary tissues. For instance, CD9, also known as tetraspanin-29, is a cell-surface glycoprotein comprised of four transmembrane regions that modulate cell adhesion and migration. CD9 has been identified as a favorable prognostic marker and predictor of metastatic potential [51]. It is highly expressed in normal colorectal epithelial tissue and forms tight connections between cells. This feature is also found in CRLM tissue. In addition, CD9 is known to be an exosome marker protein and facilitates the uptake of exosomes by recipient cells [52,53]. CD9 can therefore be used as a positive reference marker for the diagnosis of CRLM. The ALDH1A1 and GSTA1 enzymes are involved in liver metabolism. ALDH1A1 is a member of the aldehyde dehydrogenase family that participates in the oxidative pathway of alcohol metabolism in

the liver [54]. GSTA1 facilitates the conjugation of glutathione to electrophilic compounds and is also highly expressed in the liver [55]. The ALDH1A1 and GSTA1 proteins can therefore be used as negative markers for CRLM, since the metabolic features of CRC tissue are quite different and hence these markers are typically low in CRLM (**Supplementary Fig. 8**).

Moreover, our simple mass spectrometry method can be used to create a protein fingerprint of tissue samples. When combined with techniques such as deep learning, proteomics-based pathology diagnosis should facilitate the clinical diagnosis of difficult pathological sections, as well as the identification of novel markers.

## 5. Conclusions

This research used FFPE tissue slides to carry out rapid proteomic analysis and thus obtain protein fingerprints of tumor samples. Machine learning algorithms were then used to develop a predictive model for the differential diagnosis of CRLM and HCC. CD9, GSTA1, and ALDH1A1 scores were used in the prediction model, with ROC analysis showing an AUC of 0.9333. This study describes a simple method for the protein-based fingerprinting of tissues using high-throughput proteomic techniques, thereby offering a novel approach for the pathological diagnosis of CRLM and HCC.

## Availability of Data and Materials

The raw data for tissue sample proteomics was uploaded to ProteomeXchange (PXD042636). Clinical information and data used for figure illustration can be found in Supplementary Information. Further relevant data for analysis are available from the corresponding authors upon reasonable request.

## Author Contributions

Conceptualization: GY, QL and XZ; Protocol: XZ, XW and RB; Data acquisition: HL; Validation: XZ, RB; Investigation: QL, DH, XDG and GY; Manuscript preparation: XZ; Manuscript editing and revision: GY, XZ and QL; Project administration: QL, GY and XDG; Funding support: GY, XDG, and XZ. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

## Ethics Approval and Consent to Participate

This study complies with the basic principles of medical ethics of the Declaration of Helsinki, and the protocol of this experiment was approved by the Ethics Committee of the Affiliated Hospital of Jiangnan University (protocol code: NO. LS2021-048). Patient IDs and other personal information were non-publicly coded to protect patient privacy. Informed consent was obtained from all patients involved in the study. Written informed consent has been obtained from the patients or their relatives to publish this paper.


## Acknowledgment

We thank the Mass Spectrometry Center of Wuxi Medical School for their assistance with sample testing. We are grateful to the members of the Lab of Cell Glycobiology (Wuxi) and the Division of Glycobiology (Beijing) for their help with sample processing and data visualization. We would like to thank Jessica Jeffrey for the professional English language editing of this article.

## Funding

This work was supported by the National Natural Science Foundation of China (No. 32101031), China Postdoctoral Science Foundation (No. 2021M701459), the Precision Medicine Project of Wuxi Health Commission (No. jzyx04) and the Translational Medicine Research Project of Wuxi Health Commission (No. ZH202103).


## Conflict of Interest

The authors declare no conflict of interest.

## Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2901003.

## References


[1] Li X, Ramadori P, Pfister D, Seehawer M, Zender L, Heikenwalder M. The immunological and metabolic landscape in primary and metastatic liver cancer. Nature Reviews Cancer. 2021; 21: 541–557.

[2] Zhou H, Liu Z, Wang Y, Wen X, Amador EH, Yuan L, *et al*. Colorectal liver metastasis: molecular mechanism and interventional therapy. Signal Transduction and Targeted Therapy. 2022; 7: 70.

[3] Abdel-Misih SR, Bloomston M. Liver Anatomy. Surgical Clinics of North America. 2010; 90: 643–653.

[4] Hu Z, Ding J, Ma Z, Sun R, Seoane JA, Scott Shaffer J, *et al*. Quantitative evidence for early metastatic seeding in colorectal cancer. Nature Genetics. 2019; 51: 1113–1122.

[5] Pericleous S, Bhogal RH, Mavroeidis VK. The Role of Circulating Biomarkers in the Early Detection of Recurrent Colorectal Cancer Following Resection of Liver Metastases. Frontiers in Bioscience-Landmark. 2022; 27: 189.

[6] Creasy JM, Sadot E, Koerkamp BG, Chou JF, Gonen M, Kemeny NE, *et al*. Actual 10-year survival after hepatic resection of colorectal liver metastases: what factors preclude cure? Surgery. 2018; 163: 1238–1244.

[7] Villard C, Abdelrafee A, Habib M, Ndegwa N, Jorns C, Sparrelid E, *et al*. Prediction of survival in patients with colorectal liver metastases- development and validation of a prognostic score model. European Journal of Surgical Oncology. 2022; 48: 2432–2439.

[8] Biller LH, Schrag D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. JAMA. 2021; 325: 669–685.

[9] Mo S, Tang P, Luo W, Zhang L, Li Y, Hu X, *et al*. Pa-

tient-Derived Organoids from Colorectal Cancer with Paired Liver Metastasis Reveal Tumor Heterogeneity and Predict Response to Chemotherapy. Advanced Science. 2022; 9: e2204097.
[10] Cervantes A, Adam R, Roselló S, Arnold D, Normanno N, Taïeb J, *et al*. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Annals of Oncology. 2023; 34: 10–32.
[11] Kim D, Gupta B, Wong GYM. Prognostic circulating proteomic biomarkers in colorectal liver metastases. Computational and Structural Biotechnology Journal. 2023; 21: 2129–2136.
[12] Wong GYM, Diakos C, Hugh TJ, Molloy MP. Proteomic Profiling and Biomarker Discovery in Colorectal Liver Metastases. International Journal of Molecular Sciences. 2022; 23: 6091.
[13] Moldogazieva NT, Mokhosoev IM, Zavadskiy SP, Terentiev AA. Proteomic Profiling and Artificial Intelligence for Hepatocellular Carcinoma Translational Medicine. Biomedicines. 2021; 9: 159.
[14] Schoffman H, Levin Y, Itzhaki-Alfia A, Tselekovits L, Gonen L, Vainer GW, *et al*. Comparison of matched formalin-fixed paraffin embedded and fresh frozen meningioma tissue reveals bias in proteomic profiles. PROTEOMICS. 2022; 22: e2200085.
[15] Burns J, Wilding CP, Krasny L, Zhu X, Chadha M, Tam YB, *et al*. The proteomic landscape of soft tissue sarcomas. Nature Communications. 2023; 14: 3834.
[16] Coscia F, Doll S, Bech JM, Schweizer L, Mund A, Lengyel E, *et al*. A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. Journal of Pathology. 2020; 251: 100–112.
[17] Zhu Y, Weiss T, Zhang Q, Sun R, Wang B, Yi X, *et al*. High-throughput proteomic analysis of FFPE tissue samples facilitates tumor stratification. Molecular Oncology. 2019; 13: 2305–2328.
[18] Raghunathan R, Sethi MK, Zaia J. On-slide tissue digestion for mass spectrometry based glycomic and proteomic profiling. MethodsX. 2019; 6: 2329–2347.
[19] Aljawad MF, Faisal AHMA, Alqanbar MF, Wilmarth PA, Hassan BQ. Tandem mass tag-based quantitative proteomic analysis of cervical cancer. PROTEOMICS. Clinical Applications. 2023; 17: e2100105.
[20] Zecha J, Satpathy S, Kanashova T, Avanessian SC, Kane MH, Clauser KR, *et al*. TMT Labeling for the Masses: a Robust and Cost-efficient, in-solution Labeling Approach. Molecular & Cellular Proteomics. 2019; 18: 1468–1478.
[21] Yang G, Zuo C, Lin Y, Zhou X, Wen P, Zhang C, *et al*. Comprehensive proteome, phosphoproteome and kinome characterization of luminal A breast cancer. Frontiers in Oncology. 2023; 13: 1127446.
[22] Wen P, Chen J, Zuo C, Gao X, Fujita M, Yang G. Proteome and Glycoproteome Analyses Reveal the Protein N-Linked Glycosylation Specificity of STT3A and STT3B. Cells. 2022; 11: 2775.
[23] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols. 2016; 11: 2301–2319.
[24] Asleh K, Negri GL, Spencer Miko SE, Colborne S, Hughes CS, Wang XQ, *et al*. Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. Nature Communications. 2022; 13: 896.
[25] Sachs MC. plotROC: A Tool for Plotting ROC Curves. Journal of Statistical Software. 2017; 79: 2.
[26] Gu Z. Complex heatmap visualization. iMeta. 2022; 1: e43.
[27] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, *et al*. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021; 2: 100141.
[28] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, *et al*. Integrated analysis of multimodal single-cell data. Cell. 2021; 184: 3573–3587.e29.
[29] Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clinical Epigenetics. 2019; 11: 123.
[30] Zhou X, Zhai Y, Liu C, Yang G, Guo J, Li G, *et al*. Sialidase NEU1 suppresses progression of human bladder cancer cells by inhibiting fibronectin-integrin alpha5beta1 interaction and Akt signaling pathway. Cell Communication and Signaling. 2020; 18: 44.
[31] Hewitt DB, Brown ZJ, Pawlik TM. The Role of Biomarkers in the Management of Colorectal Liver Metastases. Cancers. 2022; 14: 4602.
[32] Kiritani S, Yoshimura K, Arita J, Kokudo T, Hakoda H, Tanimoto M, *et al*. A new rapid diagnostic system with ambient mass spectrometry and machine learning for colorectal liver metastasis. BMC Cancer. 2021; 21: 262.
[33] Marchione DM, Ilieva I, Devins K, Sharpe D, Pappin DJ, Garcia BA, *et al*. HYPERsol: High-Quality Data from Archival FFPE Tissue for Clinical Proteomics. Journal of Proteome Research. 2020; 19: 973–983.
[34] Davalieva K, Rusevski A, Velkov M, Noveski P, Kubelka-Sabit K, Filipovski V, *et al*. Comparative proteomics analysis of human FFPE testicular tissues reveals new candidate biomarkers for distinction among azoospermia types and subtypes. Journal of Proteomics. 2022; 267: 104686.
[35] Mitsa G, Guo Q, Goncalves C, Preston SEJ, Lacasse V, Aguilar-Mahecha A, *et al*. A Non-Hazardous Deparaffinization Protocol Enables Quantitative Proteomics of Core Needle Biopsy-Sized Formalin-Fixed and Paraffin-Embedded (FFPE) Tissue Specimens. International Journal of Molecular Sciences. 2022; 23: 4443.
[36] Barnabas GD, Goebeler V, Tsui J, Bush JW, Lange PF. ASAP horizontal line Automated Sonication-Free Acid-Assisted Proteomes horizontal line from Cells and FFPE Tissues. Analytical Chemistry. 2023; 95: 3291–3299.
[37] Mason JT. Proteomic analysis of FFPE tissue: barriers to clinical impact. Expert Review of Proteomics. 2016; 13: 801–803.
[38] Bayer M, Angenendt L, Schliemann C, Hartmann W, König S. Are formalin-fixed and paraffin-embedded tissues fit for proteomic analysis? Journal of Mass Spectrometry. 2020; 55: e4347.
[39] Sompuram SR, Vani K, Messana E, Bogen SA. A Molecular Mechanism of Formalin Fixation and Antigen Retrieval. American Journal of Clinical Pathology. 2004; 121: 190–199.
[40] Dressler FF, Schoenfeld J, Revyakina O, Vogele D, Kiefer S, Kirfel J, *et al*. Systematic evaluation and optimization of protein extraction parameters in diagnostic FFPE specimens. Clinical Proteomics. 2022; 19: 10.
[41] Föll MC, Fahrner M, Oria VO, Kühs M, Biniossek ML, Werner M, *et al*. Reproducible proteomics sample preparation for single FFPE tissue slices using acid-labile surfactant and direct trypsinization. Clinical Proteomics. 2018; 15: 11.
[42] Tanca A, Abbondio M, Pisanu S, Pagnozzi D, Uzzau S, Addis MF. Critical comparison of sample preparation strategies for shotgun proteomic analysis of formalin-fixed, paraffin-embedded samples: insights from liver tissue. Clinical Proteomics. 2014; 11: 28.
[43] van Huizen NA, Coebergh van den Braak RRJ, Doukas M, Dekker LJM, IJzermans JNM, Luider TM. Up-regulation of collagen proteins in colorectal liver metastasis compared with normal liver tissue. Journal of Biological Chemistry. 2019; 294: 281–289.
[44] Buczak K, Kirkpatrick JM, Truckenmueller F, Santinha D, Ferreira L, Roessler S, *et al*. Spatially resolved analysis of FFPE tissue proteomes by quantitative mass spectrometry. Nature Protocols. 2020; 15: 2956–2979.
[45] Sheng YH, Wong KY, Seim I, Wang R, He Y, Wu A, *et al*. MUC13 promotes the development of colitis-associated colorectal tumors via beta-catenin activity. Oncogene. 2019; 38: 7294–7310.

[46] Bajenova O, Chaika N, Tolkunova E, Davydov-Sinitsyn A, Gapon S, Thomas P, *et al*. Carcinoembryonic antigen promotes colorectal cancer progression by targeting adherens junction complexes. Experimental Cell Research. 2014; 324: 115–123.
[47] Mitsuyama Y, Shiba H, Haruki K, Fujiwara Y, Furukawa K, Iida T, *et al*. Carcinoembryonic antigen and carbohydrate antigen 19-9 are prognostic predictors of colorectal cancer with unresectable liver metastasis. Oncology Letters. 2012; 3: 767–771.
[48] Melle C, Ernst G, Schimmel B, Bleul A, von Eggeling F. Colon-derived liver metastasis, colorectal carcinoma, and hepatocellular carcinoma can be discriminated by the Ca(2$^+$)-binding proteins S100A6 and S100A11. PLoS ONE. 2008; 3: e3767.
[49] Yang Z, Zhuang L, Szatmary P, Wen L, Sun H, Lu Y, *et al*. Upregulation of Heat Shock Proteins (HSPA12a, HSP90B1, HSPA4, HSPA5 and HSPA6) in Tumour Tissues is Associated with Poor Outcomes from HBV-Related Early-Stage Hepatocellular Carcinoma. International Journal of Medical Sciences. 2015; 12: 256–263.
[50] Kong Y, Chen H, Chen M, Li Y, Li J, Liu Q, *et al*. Abnormal ECA-Binding Membrane Glycans and Galactosylated CAT and P4HB in Lesion Tissues as Potential Biomarkers for Hepatocellular Carcinoma Diagnosis. Frontiers in Oncology. 2022; 12: 855952.
[51] Lorico A, Lorico-Rappa M, Karbanová J, Corbeil D, Pizzorno G. CD9, a tetraspanin target for cancer therapy? Experimental Biology and Medicine. 2021; 246: 1121–1138.
[52] Nigri J, Leca J, Tubiana S, Finetti P, Guillaumond F, Martinez S, *et al*. CD9 mediates the uptake of extracellular vesicles from cancer-associated fibroblasts that promote pancreatic cancer cell aggressiveness. Science Signaling. 2022; 15: eabg8191.
[53] Andreu Z, Yanez-Mo M. Tetraspanins in extracellular vesicle formation and function. Frontiers in Immunology. 2014; 5: 442.
[54] Calleja LF, Yoval-Sánchez B, Hernández-Esquivel L, Gallardo-Pérez JC, Sosa-Garrocho M, Marín-Hernández Á, *et al*. Activation of ALDH1a1 by omeprazole reduces cell oxidative stress damage. The FEBS Journal. 2021; 288: 4064–4080.
[55] Mlakar V, Curtis PH, Armengol M, Ythier V, Dupanloup I, Hassine KB, *et al*. The analysis of GSTA1 promoter genetic and functional diversity of human populations. Scientific Reports. 2021; 11: 5038.