Plasma proteomic profiling of pan-cancer patients discovers biomarkers of cancers

Lin Bai[1#], Jinwen Feng[1#], Xiaoqiang Qiao[4#], Yuanyuan Qu[1,3#], Jiacheng Lyu[1#], Guojian Yang[1#], Yuanxue Zhu[2#], Hui Gao[5], Lingxiao Liao[6], Aimin Zang[2], Wencong Ding[6], Hailiang Zhang[1,3], Lingli Zhu[1], Yan Wang[7], Liang Wang[8], Xiaofang Wang[2], Yumiao Li[2], Jinghua Li[9], Xiaoping Yin[10], Guofa Zhao[2], Dan Liu[11], Xiangpeng Gao[11], Yongshi Liao[6*], Dingwei Ye[1*], Youchao Jia[2*], Chen Ding[1*]

[1]Department of Urology, Fudan University Shanghai Cancer Center, State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Science, Institute of Biomedical Sciences, and Human Phenome Institute, Fudan University, Shanghai 200433, China;

[2]Department of Medical Oncology, Affiliated Hospital of Hebei University; Hebei Key Laboratory of Cancer Radiotherapy and Chemotherapy, 212 Yuhua East Road, Baoding, Hebei 071000, P.R. China;

[3]Department of Oncology, Shanghai Medical College, Shanghai 200032, China;

[4]College of Pharmaceutical Sciences, Key Laboratory of Medicinal Chemistry and Molecular Diagnosis, Ministry of Education, Hebei University, Baoding 071002, China;

[5]Hebei Normal University of Science & Technology, Qinhuangdao, 066004, China;

[6]Department of Neurosurgery, The Affiliated Nanhua Hospital, University of South

China, Hengyang, 421002, China;

[7]Clinical Lab, Affiliated Hospital of Hebei University, Baoding, Hebei 071000, China;

[8]Department of Hematology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China;

[9]Department of hepatological surgery, Affiliated Hospital of Hebei University, 212 Yuhua East Road, Baoding, Hebei 07100, P.R. China;

[10]CT/MRI room, Affiliated Hospital of Hebei University, 212 Yuhua East Road, Baoding, Hebei 071000, P.R. China;

[11]College of Clinical Medicine, Hebei University, Baoding 071002, China; Department of Medical Oncology, Affiliated Hospital of Hebei University, Baoding 071000, China;

[*]To whom correspondence should be addressed. Email: liaoys66@163.com (Y. S. L), dwyeli@163.com (D. W. Y.), youchaojia@163.com (Y. C. J.), chend@fudan.edu.cn (C. D.)

[#]These authors contribute equally

Abstract

Circulatory system proteins play a central role in human physiology; therefore, profiling plasma proteins has been explored widely as a robust and dynamic tool in studying diseases. Here, we investigated the proteomic profile of plasma in 1,118 pan-cancer patients. The integrative analysis uncovered diverse molecular characteristics of different cancer types and bridged the plausible connection with the clinical features, including tumor stages. The findings demonstrate that the balance between lipid and glucose metabolisms is important in regulating immune infiltration. The pre- and post-surgery comparison of the plasma proteome indicated that tumor-induced angiogenesis is overexpressed in the tumor, and the differences between the proteome patterns could monitor post-surgery therapeutic effects. Finally, we developed a panel of tumor-type-specific proteins to classify tumor types with >95% sensitivity/specificity. Collectively, this study portrayed the plasma proteome landscape of human cancers and screened reliable biomarkers that could assist in diagnostic and drug discovery efforts.

Keywords: Plasma proteins, proteomics, pan-cancer, biomarker

Background

Omics analysis has markedly changed the clinical management of several malignancies[1, 2]. Several international consortia, including The Cancer Genome Atlas (TCGA),

Memorial Sloan Kettering Cancer Center (MSKCC), and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), have performed comprehensive genomic and proteomic analysis of multiple tumor types[3-5]. Tumor gene expression data have revealed distinct tumor subtypes and uncovered expression patterns associated with clinical outcomes. Furthermore, several landmark studies have also demonstrated that gene expression data could provide valuable information about tumor characteristics[6-8]. These efforts have dramatically refined our understanding of cancer biology, including the mechanism of carcinogenesis, development, and metastasis. However, these studies mainly focus on the tumor tissues. The molecular characteristics of the noninvasive biopsies, such as plasma, remain poorly understood in most human cancers, which impede the discovery of the cancer biomarkers.

Plasma is the predominant sample used for diagnostic analyses in clinical practice and is available in bio-banks from thousands of clinical studies. Even though the plasma serves as the major biomarker resource for various diseases, tumor diagnosis is still challenging. Recently, several studies have demonstrated the potential of noninvasive blood-based liquid biopsy and circulating cell-free DNA (cfDNA) analysis in cancer diagnosis and management[9-11]. However, the broad insights gleaned from genomic information derived from cfDNA introduce additional complexities compared with tissue-based genome profiling. Moreover, the detectability of circulating tumor DNA (ctDNA) is limited by tumor size, the trace ctDNA amount in the blood, and the

technological platform[12].

Plasma proteins account for the major functional component in the plasma and play key roles in various biological processes, including signaling, transport, growth, repair, and defense against infections. These proteins are frequently dysregulated in disease and are important drug targets. Although the plasma biomarkers, such as carcinoembryonic antigen (CEA), carbohydrate antigen 199, and alpha fetoprotein, were widely used to indicate several cancers, including colon cancer, pancreatic cancer, and liver cancer, the current blood tumor protein biomarker pools are far from meeting the clinical requirement in screening human cancers. Furthermore, mass spectrometry (MS)-based plasma proteomics is extremely challenging for several reasons, prominently attributed to the extremely large dynamic range of protein abundances. We recently developed a rapid and robust ‘‘plasma proteome profiling’’ pipeline following a data-independent acquisition (DIA) based MS strategy. The approach established a large-scale DIA reference library that can quantitatively identify over 2,000 proteins in a single MS run in one hour.

In this study, we report the proteomics characterization of a large cohort of tumor plasmas, to date, by profiling the plasma of 1,118 treatment-naive pan-tumor patients. Our study reports a prediction model to distinguish the tumor types, unravels the featured carcinogenesis pathways of different tumor types and bridges the plausible connection with the clinical features. The explored data represent an exceptional resource for further

biological, diagnostic, and drug discovery efforts.

Results

Proteomics analyses of pan-cancer specimens

We collected 1,118 plasma samples from treatment-naive patients with tumor in China, which included lung cancer (LC, n=233), malignant lymphoma (ML, n=139), renal carcinoma (RCC, n=130), bladder cancer (BLCA, n=121), breast carcinoma (BRCA, n=101), colorectal cancer (CRCA, n=65), stomach adenocarcinoma (STAD, n=49), testicular germ cell tumors (TGCT, n=46), esophageal carcinoma (ESCA, n=44), ovarian cancer (OVCA, n=25), cervical carcinoma (CECA, n=21), liver hepatocellular carcinoma (LIHC, n=21), cholangiocarcinoma (CHOL, n=18), sarcoma (SARC, n=17), pancreatic adenocarcinoma (PAAD, n=11), thyroid carcinoma (THCA, n=11), head and neck carcinoma (HNCA, n=10) and others (tumor types under count 10, n=56) (**Table S1**). Further, to build up the baseline of the plasma proteome of the healthy population for the reference map, we used 200 non-tumor-derived plasma cases (NTDP) (**Figures 1A and S1A; Table S1**). Clinicopathological indicators, including tumor type, tumor node metastasis (TNM) stage, age, gender, and 93 biochemical indicators, such as platelet (PLT) level, red blood cell count (RBC), and hemoglobin level (HGB), were evaluated and summarized for all cases (**Figures 1B and S1B; Table S1**).

The pan-cancer plasma proteome analysis was performed by high-resolution liquid chromatography-mass spectrometry (LC-MS/MS) using an orbitrap Fusion Lumos mass spectrometer. The stability of the mass spectrometry (MS) platform was assessed by quality control (QC) runs. A spearman's correlation coefficient was calculated for 50 quality control (QC) runs using HEK293T cell samples (**Figure S1C**). The average correlation coefficient of the QC samples was 0.91 (range, 0.84-0.97). Processed data are shown in **Table S1** (raw data available via Data Resources in **Methods**). The number of proteins (at a peptide- and protein-level false discovery rate (FDR) of 1%) in each sample ranged from 1,155 to 4,247 (**Figure 1C**) and was significantly higher in tumor sample plasma (median, 2,381) than that in NTDP (median, 2,356) (t-test, p-value < 0.0001, **Figure S1D**). In total, 10,946 proteins were detected in the 1,318 samples, among which 9,308 proteins were common between tumor- and NTDP samples, whereas 1,535 and 103 proteins were specific in tumor-derived plasma (TDP) and NTDP, respectively (**Figure 1D; Table S1**). On average, about 2,100 proteins were measured in each tumor type (**Figure 1C; Table S1**), and 4,267 proteins detected at least in one sample were present in all 19 tumor types (**Figure 1E**). In this study, the individual plasma proteomes were characterized by their relative abundance, not by the simple presence or absence of proteins. Proteome quantification was conducted using the intensity-based absolute quantification (iBAQ) algorithm, followed by the fraction of total (FOT) normalization as reported previously. The dynamic range of proteins detected spanned eight orders of

magnitude (**Figure 1F**). Principal component analysis (PCA) revealed a clear distinction among 18 tumor-derived plasma proteomes, indicating that protein variation between tumor types exceeds the variation between individuals (**Figure 1G**).

Searching the identified proteins (10,946) in several databases (**Table S1**) revealed several functional groups, including 3,237 drug targets, 406 transcription factors (TFs), 45 cytokines, 2,322 enzymes, 356 kinases, 66 growth factors, 32 hormone activities, and 92 G protein receptors (**Figures 1H, 1I, S1C, and S1D**). The distributions of these proteins in 18 tumor types are shown in **Figures S1E and S1F**. Collectively, these data, indicating sufficient coverage for analysis of biological processes, represent a comprehensive plasma proteomic resource for the tumor study.

Proteomic classification revealed heterogeneity of pan-cancer

To investigate the intrinsic structure of the tumor plasma proteomic data, non-negative matrix factorization (NMF)-based unsupervised clustering was performed on the most variable proteins (proteins with the top 3,000 coefficients of variation (CV)). The six NMF clusters (C1–C6) were identified (**Figures S2A and S2B**), comprising 96, 490, 66, 80, 133, and 253 cases.

To assess the distribution of tumor type in each cluster, we calculated the tumor type enrichment score (TTES) for each cluster. We found C1 was enriched with CRCA and STAD (TTES = 1.61, 1.43), C2 was distinguished by BRCA, OVCA (TTES = 1.47, 1.55),

C3 was classified by LIHC, BLCA (TTES = 2.42, 1.68), C4 was enriched with HNCA, TGCT, LC, CECA (TTES = 2.80, 1.82, 1.68, 2.00), C5 mainly consisted of CECA (TTES = 1.60), and C6 was characterized by RCC and BLCA (TTES = 3.40, 3.18). According to the TTES of the tumor types, we termed C1 to C6 as Gastrointestinal-rich, BRO-rich, Urinary I-rich, Mixed, Cervical-rich, and Urinary II-rich (**Figure 2A; Table S2**).

To further explore the biology associated with the proteomic taxonomy, we performed consensus clustering using the ConsensusClusterPlus (CCP) package based on differentially expressed proteins (DEPs) (Kruskal-Wallis, adjusted p-value < 0.05, **Figure 2A**) and identified nine protein groups (**Figure S2C**) belong to the above six clusters. Further, we carried out over-representation analysis (ORA) based on these protein groups. The Gastrointestinal-rich cluster was primarily associated with acute inflammatory response (FDR = 2.12E-10) and complement and coagulation cascades (FDR = 7.31E-38). The BRO-rich cluster was demonstrated as MYC targets (FDR = 4.62E-4) and associated with translational processes such as RNA splicing (FDR = 1.25E-9) and translational initiation (FDR = 3.18E-6). The Urinary I-rich cluster was distinguished by high level of heme metabolism (FDR = 5.67E-4) and hydrogen peroxide metabolic process (HPMP; FDR = 2.28E-11). The Cervical-rich cluster showed upregulation of estrogen signaling (FDR = 1.27E-9), extracellular matrix organization (FDR = 0.030), and cornification (FDR = 1.56E-7). The Urinary II-rich cluster was outstood by pancreatic secretion (FDR = 0.012), activation of JUNK activity (FDR =

6.23E-3), and carbohydrate derivative catabolic process (FDR = 0.025). The mixed cluster showed higher level of positive regulation of NF-κB activity (FDR = 2.41E-5) and toll-like receptor signaling pathway (FDR = 2.74E-3). Pathway scores based on global proteomics data which obtained by single-sample gene set variation analysis (ssGSEA)[13] also demonstrated similar results. Taken together, we inferred that the clusters and the pathway enrichment analysis highlighted intrinsic heterogeneity in oncogenic signaling across different tumor types.

We then accessed the enrichment of the clinical annotations in different clusters and found the significant connections between particular cluster types and clinical information. As shown in **Figure 2B**, the Gastrointestinal-rich cluster was dominated in TNM stage IV, indicating poor clinical outcomes. Furthermore, STAD and CRCA, the highly enriched tumor types in this cluster, were all annotated as TNM stage IV types (Chi-square p-value = 0.05, **Figures 2C and S2D**). Consistently, the TNM stage IV of STAD and CRCA were remarkably enriched in Gastrointestinal-rich cluster than in other clusters (**Figure 2D**, Chi square p-value = 0.03), indicating that the Gastrointestinal-rich cluster represented a terminal stage of the gastrointestinal tumor patients.

Gene set enrichment analysis (GSEA)[14] for GO biological processes using the signal to noise value between the STAD/CRCA in the Gastrointestinal-rich cluster and other clusters revealed acute phase response (APR) was upregulated in the Gastrointestinal-rich cluster (Normalized enrichment score = 2.39, FDR = 0, **Figure 2E**). Notably, several

APR markers such as CRP, SAA1, and SERPINA1 had the highest expression level in the Gastrointestinal-rich cluster (Ranksums p-value<0.01, **Figures 2F and 2G**). CRP is the most important acute phase reactant. In an attempt to explore the effects of CRP in STAD/CRCA, we performed pair-wise Spearman correlation on CRP and clinical annotations, which revealed a significantly higher correlation with neutrophil percentage (NE%) (**Figure 2H**). Moreover, NE% was significantly higher in the Gastrointestinal-rich cluster than in other clusters (Ranksums p-value<0.01, **Figure 2I**). To evaluate the role of neutrophils in tumor metastasis, we applied the Spearman correlation algorithm to NE% and the proteome data in all samples and discovered neutrophil-related metastasis proteins[15-18] (**Figure 2J**). The results revealed that the metastasis signature proteins S100A8 and S100A9 were significantly correlated with NE% (Spearman rho>0.3). Besides, we also found that the S100A8/S100A9 were overrepresented in the Gastrointestinal-rich cluster than in other clusters (**Figures 2K, 2L, and S2E**), indicating S100A8/S100A9 play key roles in the metastasis of the STAD/CRCA. These observations suggested that the CRP-promoted neutrophils recruitment can contribute to metastasis in the Gastrointestinal-rich cluster by increasing the expression level of S100A8/S100A9 and leading to a worse prognosis (**Figure 2M**).

As shown in **Figure S3A**, the Urinary I-rich cluster was characterized by erythrocyte-related clinical annotations. HGB was increased in this cluster (Kruskal-Wallis, p<0.0001, **Figure 3A**). In this cluster, BLCA, but not LIHC, showed a

significantly high level of HGB than other tumor types (**Figures 3B and S3B**), which was higher than that in other clusters (Ranksums p-value<0.01, **Figure 3C**). Interestingly, all BLCA patients classified in the Urinary I-rich cluster were male (**Figures 3D and S2D**), and the HGB level was the highest in the male population in the Urinary I-rich cluster than the male population in the other clusters and the female populations (Ranksums p-value<0.05) (**Figure 3E**).

The GSEA analysis on the BLCA male population between Urinary I-rich and other clusters showed that heme metabolism and HPMP were enriched in the Urinary I-rich cluster than in other clusters (**Figure 3F**). Major proteins participating in these processes, such as BLVRB, CA1, CAST, and PRDX2, were upregulated in the Urinary I-rich cluster (**Figure 3G**). Further, we calculated Spearman correlation based on the HGB clinical values and process ssGSEA scores to investigate the relation between HGB and the above two biological processes. The results indicated that HGB was significantly correlated with heme metabolism (Spearman rho = 0.419, p-value = 2.18E-5), and the latter was strongly correlated with HPMP (Spearman rho = 0.592, p-value = 8.93E-11, **Figures 3H and 3I**), which was consistent with a previous report[19, 20]. Here, we estimated Multi-Gene Proliferation Scores (MGPS) to assess the influence of HPMP on cell proliferation, which revealed a negative association between HPMP and MGPS (Spearman rho = −0.314, p-value = 1.47E-3, **Figure 3J**) and BLCA patients in Urinary-I rich cluster had a lower MGPS level than in other clusters (Ranksums p-value = 0.03, **Figure 3K**). Consistently,

we found more early T stages of BLCA patients in the Urinary I-rich cluster (**Figure S3C**). Subsequently, to investigate the relation between the HGB and the prognosis of the BLCA, we referred to TCGA BLCA data for survival analysis. Significantly, the level of HBA2 was positively related to better prognosis in males but not in females (**Figures S3D and S3E**). These findings uncovered the linkage between HGB and BLCA (**Figure 3L**) and suggested that HGB could be considered a prognostic biomarker for male BLCA patients.

The Urinary II-rich cluster, which has the highest creatinine (CREA) level, consisted of the highest number of Urinary cancers (Kruskal Willis, p = 4.05E-12, **Figures 3M and 2A**). The elevated level of CREA, a break-down product of creatine, signifies impaired kidney function or kidney disease. To investigate how CREA affects the carcinoma process, we used correlation tests between CREA and Gene Set Variation Analysis (GSVA) scores. As a result, in Hallmark gene sets, we identified pancreas beta cell pathway ranked the top showing strong correlation with CREA (Spearman rho = 0.197, FDR = 2.00E-5, **Figure 3N**) and was overrepresented in the Urinary II-rich cluster (**Figure S3I**), which is consistent with the report that high level of creatine induced insulin secretion[21]. In addition, insulin receptor signaling pathway (INSR, PDK4, RGN, SLC1A2, etc.) that showed higher association with pancreas beta cell pathway (Spearman rho = 0.199, p-value = 2.028E-11, **Figure 3O**) was activated in this cluster (Kruskal-Wallis p-value = 2.038E-25, **Figure S3F**). Further analysis revealed that

glucose import (Spearman rho = 0.330, FDR = 1.10E-27) / glucose metabolic process (Spearman rho = 0.184, FDR = 6.37E-9) have strong correlation with the insulin receptor signaling pathway (**Figure 3P**) and were activated in Urinary II-rich cluster (**Figure S3I**). Insulin-regulated facilitative glucose transporter, GLUT4, HNF1A, and ASPSCR1 were all elevated in the Urinary II-rich cluster (FDR<0.01, **Figure 3Q**), especially in RCC. In addition, we found lymphocyte count (LY) was elevated in this cluster (Kruskal-Wallis p-value = 3.513E-10, **Figure S3G**). Therefore, we hypothesized an imbalance between glucose metabolic process and immune infiltration. To confirm this hypothesis, we observed the correlation between the glucose metabolic process and LY (Spearman rho = 0.241, p = 7.659E-14, **Figure S3H**). Above all, these results highlighted the abnormal level of CREA that leads to the interplay between metabolic reprogramming and immune function linked by insulin signaling (**Figure 3R**).

Plasma proteome based hierarchical clustering of pan-cancer

To investigate TDP proteome features in different physiological systems, we classified 15 tumor types into six physiological systems, including excretory system (BRCA, CHOL, LIHC, PAAD, and THCA), reproductive system (CECA, OVCA, and TGCT), urinary system (BLCA and RCC), digestive system (CRCA, ESCA, and STAD), immune system (ML) and respiratory system (LC). Consequently, a total of 1,323 (Kruskal-Willis test, p-value < 0.05, the expression of the proteins in a certain system was >2 folds higher

than the other, consistently) system-specific plasma proteins were identified, ranging from 94 TDPs in the respiratory system to 347 TDPs in the digestive system. (**Figure 4A; Table S3**). To reveal the functional characteristics of the system-specific plasma proteins, we performed the Kyoto Encyclopedia of Genes and Genomes (KEGG) function enrichment analysis. The results demonstrated that the excretory system-enriched plasma proteins were significantly enriched in pathways (p-value < 0.05), including mitochondrial translation (such as MRPS35, GFM2, and MRPL1), energy metabolism (such as EP300, TFAM, and RXRA) and endocytic vesicle (such as SNX3, PLD4 and PLD1). The plasma proteins highly expressed in the digestive system were mainly involved in pathways related to linoleate metabolism (such as PLA2G4A, CYP3A5, and UGT1A7), ribosome (such as MRPS7, RPL10, and RPL18), androgen and estrogen biosynthesis and metabolism (such as CYP3A5 and COPE) and proteoglycan biosynthesis (EXT1 and COPE) (**Figure 4B**). Collectively, we found the specific system plasma proteins were significantly enriched in the predominant biological processes of the corresponding systems.

Intriguingly, the specific system plasma proteins were also correlated with the TNM stage. We identified **60** specific system plasma proteins based on their significant association with the TNM stage (Kruskal-Wallis test, p-value < 0.05, correlation > 0.2). For example, IGFBP5, KRT13, KRT15, ORM2, and PSIP1 were the digestive system-specific plasma proteins, whereas VCAM1 and HLA-B were significantly upregulated

in the immune system. There were 11 (C4BPA, CFB, C9, SERPINA1, SERPINA3, CRP, HSPB1, ITH3, PSMB10, RCN1, and CP) proteins, which were correlated with the TNM stages in the urinary system (**Figure 4C**). In addition, there were 20, 12, and 10 proteins, which were correlated with the TNM stages in the excretory system, respiratory system and reproductive system, respectively (**Figures S4A and S4C**). These findings suggest that identifying specific system plasma proteins associated with TNM staging could disclose the mechanism in the tumor progression and nominate the biomarker to identify the tumor stages.

The hierarchical clustering showed that patient samples from the same physiological system tended to cluster together, indicating more similarity in tissue-originated tumor type (**Figure 4D**). Consequently, we identified four clusters (subtypes) based on the correlation between the 19 tumor types. A total of 1,124 (Kruskal-Willis, p-value < 0.05, the expression of the proteins in the certain subtypes was > 2-folds higher than the other, consistently) most variable proteins, ranging from a minimum of 205 proteins in cluster 2 to a maximum of 455 proteins in cluster 4 were identified. The clusters revealed a clear co-clustering among the samples from the GI tract (colorectal, stomach, and esophageal) and the urinary organs (renal and bladder) (**Figure 4D, Table S3**). Furthermore, KEGG analysis of these 1,124 differentially expressed proteins revealed that cluster 1 included ML, CRCA, BRCA, LC, ESCA, STAD, and OVCA and was associated with fatty acid metabolism (ACOT8, CHKB, ACSF3, and ALOX5AP), cell cycle (HDAC1, CDC23,

NUDC, and WAPL) and membrane trafficking (MADD, RAB31, ATAM2, and LDLR). Cluster 2 that included RCC, BLCA and CECA, was characterized by the highest level of ErbB signaling pathway (AKT1, PAK2, SHC1, and ERBB4), PI3K/AKT signaling (FGFR2, GSK3A, and TSC2), and neutrophil degranulation (ALDH3B1, NRAS, CYBB, and LAMTOR1). Cluster 3 comprising LIHC and SARC was enriched in pathways of cholesterol biosynthesis (FDPS and SQLE) and catabolism of glucuronate to xylulose-5-phosphate (SORD and DCXR). Cluster 4 included HNCA, THCA, TGCT, PAAD, and CHOL and was connected with VEGFR signaling (PIK3R1, CBL, and FLT1) and glucose metabolism (PFKFB1, SLC37A4, and POM121) (**Figure 4E**).

To further elucidate the underlying differential characteristic of these four clusters, we investigated their association with clinical indices. Serum total bile acid (TBA) level, a sensitive marker of liver function, is used to diagnose and monitor various liver diseases[22]. In our dataset, the TBA level was overrepresented in cluster 3 (Kruskal Willis P-value = 2.2e-06, **Figure S5A**), and the LIHC in this cluster showed the highest TBA level (t-test P-value = 0.0009, **Figure S5B**).

To illustrate the mechanism of the high level of TBA in LIHC, we performed a Spearman correlation algorithm on TBA and global proteome data and ORA on significant proteins (p-value < 0.05, Spearman rho > 0). The results indicated that proteins that participate in bile acid metabolism (FABP6, OSBPL2) and metabolism of lipid (FDPS, PCYT1B, ACSL6, LPGAT1, PLA2G15, SPTLC2) process showed a

stronger correlation (**Figure S5C**) with the TBA. Furthermore, FDPS, a key intermediate in cholesterol and sterol biosynthesis, which facilitate TBA biosynthesis[22], displayed a high expression level in LIHC (**Figure S5D**). To confirm its function in LIHC, we examined FDPS expression in the TCGA LIHC cohort and observed an upregulated level in LIHC tissue than adjacent tissues (**Figure S5E**). From these findings, we hypothesized that the activated metabolism of lipid could lead to TBA biosynthesis disorders by upregulating the FDPS expression level (**Figure S5F**).

Cluster 4 consists of a mixture of HNCA, SARC, TGCT, PAAD and CHOL. We found that the patients in cluster 4 showed an elevated level of albumin (ALB) (Kruskal-Wallis p-value = 2.803E-7, **Figure 4F**), and TGCT showed an increased ALB level than other tumor types in this cluster (Kruskal-Wallis p-value < 0.0001, **Figure 4G**). Serum ALB levels have been shown to have prognostic significance in cancers[23]. To illustrate the reason for higher-level ALB in TGCT, we firstly examined the pathways related to ALB biosynthesis. The branched-chain amino acid metabolism has been reported to promote ALB biosynthesis[24]; however, we did not find the upregulation of this pathway in the TGCT, indicating the high level of ALB might not be because of the superfluous biosynthesis (**Figure S5H**). Further, to unravel the possibilities of another speculation that the blood-testis barrier (BTB) can influence the transport of ALB and tight junction plays an important role in BTB formation[25], we performed the Spearman correlation analysis between ALB levels and pathway GSVA scores. The results demonstrated a

strong correlation of the adhesion-related pathways with ALB (**Figure 4H**), and the cell junction assembly was consistently activated in TGCT (Kruskal-Wallis p-value = 1.7E-7, **Figure 4I**). RAC signaling was reported as the upstream of cell junction assembly[26]. We also found a significant positive correlation between cell junction assembly and RAC protein signal transduction in our cohort (r = 0.24, p-value = 2.7E-16, **Figure 4J**). Moreover, the cell junction assembly related proteins, including TRIP6, ARVCF, S100A10, STMN1, RAB29, NEBL, CNN2 were also found to be strongly correlated with RAC protein signal transduction (**Figure 4K**), and these proteins showed the highest level in TGCT than in other tumor types in cluster 4 (**Figure 4L and S5G**). Collectively, we inferred that the RAC signaling might hyper-stimulate the BTB formation by increasing the activation of cell junction assembly and thus, preventing the ALB transport in TGCT (**Figure 4M**).

Identification of specific tumor-derived plasma proteins

Several blood proteins, such as CEA (colorectal cancer and pancreatic carcinoma) and CA125 (ovarian cancer), have been reported for tumor detection. To explore the role of these conventional blood markers, seven blood markers (CEA, NSE, CA153, CA125, PAP, SCCA1, and SCCA2) were examined in our cohort (**Figure 5A**). Specifically, compared to others tumor types, OVCA had the highest expression of CA125, which was consistent with the previous study[27]. The most commonly increased plasma protein

markers associated with cancer were CEA (BLCA and PAAD), NSE (RCC, BLCA, ESCA, CECA, and PAAD), CA153 (BLCA, STAD, OVCA, and CECA), PSAP (ML, BLCA, OVCA, LIHC, and CHOL) and SCCA1 and SCCA2 (LC, RCC, and THCA) (**Figure 5B**). However, we found that some biomarkers were also highly overrepresented in other cancer types besides the targeting cancer type, such as CA153 in BLCA and STAD. These results indicated the limitation of a single biomarker in identifying specific tumor types. Therefore, we focused on expanding the spectrum of novel blood-derived biomarkers.

To identify the TDP markers, we stratified the proteins into two categories: non-ubiquitous proteins (non-ubiPros, 3,107, accounting for 58.8% of the identified proteins), which were highly expressed in only a few tumor types with a transformed median expression value of < 0.5; ubiquitous proteins (ubiPros, 2,179, 41.2%), which were expressed in a wide variety of tumors with a transformed median expression value of > 0.5 (**Figure S6A**). We identified 3,107 tumor-specific plasma proteins, ranging from 49 proteins in LC to 330 proteins in CHOL. We then performed KEGG analysis of the tumor-specific plasma proteins and found consistency between the specific plasma proteins with the relative feature of tumor type (**Figure 5C**). For example, HIF signaling pathways (MAP2K1, EIF4E) and peroxisome (ABCD1, ECH1, and GNPAT) were enriched in RCC tumors, indicating a significant Warburg effect in RCC tumor, which enables tumor cells to rapidly proliferate and survive under nutrient depletion and

hypoxia[28].

To further connect specific co-expression of the protein groups with different tumor types, we performed weight gene correlation network analysis (WGCNA)[29] (**Figure 5D)**. Separating the proteomic profiles to eigengene modules identified a group of modules positively associated with different tumor types (p-value < 0.05, r > 0.92). WGCNA identified 18 protein modules, and the number of proteins in different modules ranged from 49 proteins in module grey60 related to LC to 330 proteins in module turquoise related to CHOL. Moreover, the findings demonstrated an extensive connection between modules with tumor types. For instance, module purple (MAP3K9, NUDCD2, ATL1, ERBB4, etc.) was highly correlated to BLCA (p-value < 1.1E10, r = 0.95). The proteins AQP2, CLCA4, RXRA, MRPL13 comprising module red were strongly associated with PAAD. Further, based on the cut-off criteria (GS > 0.2, correlation > 0.8), we identified the proteins with high connectivity in the significant module as hub proteins, which were also found to be commonly highly expressed in the corresponding specific tumor type. For example, phosphodiesterase [ENPP3], which has been shown to be involved in tumor cell migration[30, 31], was identified as a hub protein and highly expressed in RCC. Moreover, SLC27A1, one of the three members of the fatty acid transport protein family, was overrepresented in LIHC. As fatty acids are fundamentally required for tumor cell proliferation to provide new phospholipids for plasma membranes, the increase of the SLC27A1 might suggest a demand for the uptake of exogenous fatty acids. Consequently,

the inhibition of exogenous fatty acids uptake might provide novel therapeutic approaches to treat this pernicious tumor type. We classified the top 10 functional hub proteins in the modules with the highest connectivity as candidate markers for further analysis **(**Top 10 proteins in **Figure 5E**; complete list in **Table S4)**.

Interestingly, we found some of these hub proteins were associated with the TNM stage. For example, in RCC samples, there were 20 potential biomarkers (Kruskal−Wallis test, p-value < 0.05) positively associated with the TNM stage (**Figure 5F**). Gene Ontology (GO) analysis demonstrated that most of these proteins were involved in immune response (SERPINA1/3, ORM1/2, CFI, C2, C9, and LRG1). We next identified 12 potential biomarkers negatively associated with the TNM stage in RCC samples, which were mostly involved in blood microparticles (AFM, ITIH2, and GSN), endomembrane system (ALB, LUM and SERPINA4), and plasma lipoprotein particle (APOM, APOA1) (**Figure 5G**). These findings suggest that identifying the potential markers associated with the TNM staging could guide physicians to select appropriate therapeutic schedules.

We reasoned that ideal biomarkers should be overexpressed in the corresponding tumor tissue (upregulated in the tumor tissue sample) and released into the blood (upregulated in the plasma). To this end, we performed tissue proteome screening of eight tumor types (ESCA, BLCA, STAD, ML, CRCA, BRCA, RCC, and PAAD). For each tumor type, we collected formalin-fixed paraffin-embedded (FFPE) samples of tumor

tissues and the matched adjacent tissues from 10 patients (**Figure S6B**). The FFPE tissue proteome identified 16,913 proteins in total, ranging from a minimum of 10,375 proteins in BLCA tumors to a maximum of 10,577 proteins in ML tumors. To identify tumor-specific proteins, we first identified tumor-specific tissue proteins by comparing the tumor tissue proteome and matched adjacent tissue proteome (fold change > 2, adjusted p-value < 0.05). Meanwhile, we also filtered tumor-specific plasma proteins by processing the same comparison between TDP and NTDP proteomes (fold change > 2, adjusted p-value < 0.05). Proteins with potential biomarker value and biological relevance in tissue proteome overlapped with plasma proteome are shown in **Figures 5H, S7A, 7B, S8A, and 8D**. The proteins overlapped in the eight tumor types are listed in **Table S4**. Consequently, collating these data, 585 potential biomarkers in eight tumor types were identified-for instance, 153 potential biomarkers in LC were identified (**Figure 5I**). Of note, these proteins are involved in several processes, including cell cycle, spliceosome, and transcription-coupled nucleotide excision repair (**Figure 5J**). In addition to the potential plasma biomarkers overrepresented in tumor samples, we also identified proteins that were exclusively detected in TDP and generated a list of proteins detected in > 50% of TDP samples but not in NTDP samples. Using a cut-off of > 5-fold upregulation of proteins (t-test, p-value < 0.05) in the tumor tissues than the matched adjacent tissue proteome, 254 core marker proteins were determined totally. For example, 94 core markers out of 153 potential markers were identified in the LC, including EIF3D,

MCM6, and RFC5 (**Figure 5K; Table S4)**. Notably, among the 48 proteins specifically expressed in PAAD, only 12 proteins (COL8A1, LIG1, PLOD2, etc.) were exclusive to PAAD derived plasma proteins. Similarly, 19 proteins in BLCA (EPHA1, PPFIA2, TMEM65, etc.), 9 proteins in BRCA (HMGN1, PNN, TCEA1, etc.), 8 proteins in RCC (ANAPC1, FYB1, MCM6, etc.), 23 proteins in CRCA (ANAPC1, CDH11, CNN2, XPO4, etc.), 75 proteins in ESCA (ELANE, PDCD11, SLFN5, etc.), and 14 proteins in ML (FANCI, ANAPC1, RFC5, etc.) were determined as the core markers, respectively. Collectively, these core biomarkers indicated heterogeneity among different tumor types and could be used to discriminate cancer types.

Immune infiltration in pan-cancer tumors

To gain insight into immune features in pan-cancer, we performed cell type deconvolution analysis using xCell[32] based on proteome data to infer the relative abundance of different cell types in the tumor microenvironment **(Figure 6A; Table S5**) and used ESTIMATE[33] to infer the immune score of tumor samples (**Figures 6A and 6C**). Consensus clustering based on the inferred cell proportion identified five sets of tumors with the dominant presence of specific cell types and pathways, defined as *immune clusters (ICs) 1-5*, comprising 236, 148, 251, 250, and 233 cases, respectively (**Figures 6A, S9A, and S9B; Table S5**). Comparing this with tumor types, we observed

lower immune infiltration in ML, CRCA and LC and a higher immune infiltration in RCC and BLCA (**Figures 6A and 6C**).

The *IC1* group, containing a mixture of BRCA (n=100), ESCA (n=44), HNCA (n=10), LIHC (n=21), TGCT (n=46), and THCA (n=11) samples, was characterized by the presence of multiple types of immune cells, including mast cells, osteoblast, and $CD4^{+}$T-cells (**Figures 6A and 6B**). *IC2* group, comprising a mixture of CHOL (n=18), CECA (n=21), OVCA (n=25), PAAD (n=11), SARC (n=17), and other (n=56) samples, exhibited enrichment of Tregs cells. These two groups shared several pathway features, including upregulation of inositol phosphatase metabolism and sphingolipid biosynthesis (Kruskal−Wallis test, p-value < 0.05) (**Figures 6A and 6B**).

The *IC3* group was characterized by a high degree of various types of immune cells such as $CD8^{+}$ Tem, B-cells and $CD8^{+}$ T-cells, containing BLCA (n=121) and RCC (n=130) plasma samples (**Figures 6A and 6B**). Moreover, compared with other tumors, the *IC3* cluster displayed the highest immune score and upregulation of MAPK activation, PI3K-AKT signaling in cancer and downregulation of lipids metabolism (e.g., SUMF2, ACOX2, PLPP3, CD36) (**Figures 6A, 6C, and S9C**). This is consistent with recent reports that tumors with high levels of immune infiltration are characterized by lower action of lipids metabolism[34] (Spearman's correlation, r = −0.26, p-value < 2.20E-16) (**Figure 6D**). The highest overlap (67%) was observed between the *IC3* group and the Urinary II-rich cluster (**Figure 6E**), and the results agree with those shown in **Figures**

S3G and S3H. Furthermore, in the *IC3* group, we found a higher activation of glucose metabolism (e.g., GSK3B, PGP, PHKG2) (Kruskal−Wallis test, p-value < 0.05) (**Figure 6G**), and a significant positive correlation between the pathway scores of glucose metabolism and scores of immune infiltration (Spearman's correlation, r = 0.14, p = 1.70E-06) (**Figure 6F**). These results revealed the facilitative and inhibitory effects of glucose metabolism and lipid metabolism on immune infiltration, respectively (**Figure 6H**).

The *IC4* cluster, containing a mixture of CRCA (n=63), ML (n=137) and STAD (n=49) samples (**Figures 6A and 6B**), was characterized by a high degree enrichment of epithelial cells, endothelial cells, and CLP. Proteomic analysis showed upregulated pathways involved in mitochondrial fatty acid beta-oxidation (e.g., IDH3G, HSCB), G protein activation (e.g., ABCC8, KCNK2) and apoptosis (e.g., BAX, E2F1).

IC5, as the group of LC enrichment (n=233), was characterized by upregulation of peroxisomal lipid metabolism, bile acid and bile salt metabolism, and steroid hormone biosynthesis, which showed a high degree enrichment of megakaryocytes and astrocytes (**Figure 6A**). In patients with LC, venous thromboembolism (VTE) is a recognized complication and one of the main causes of death[35]. Importantly, we found that megakaryocytes, the precursors of platelets, showed the most enrichment levels in *IC5* (**Figure 6I**) and exhibited a positively significant correlation with platelet-associated proteins (e.g., SELP, CALM) (**Figures 6J and 6K**). Identically, *IC5* exhibited the highest

platelet (PLT) level (**Figure 6L**) than other *ICs*. Corresponding with the previous studies, we found that PLT was positively correlated with Fibrinogen (FIB) and elevated level of FIB in *IC5* consistently (**Figures 6M and 6N**). These findings are consistent with those of the recent reports that show under abnormal FIB, the blood will continue to coagulate, leading to thrombosis[36]. Taken together, the megakaryocyte level associated with diverse clinical features could reveal the molecular mechanism influenced on thrombus production in LC plasma (**Figure 6O**).

Short-term changes of the plasma proteome after surgery

Surgery, the operation to remove part of the body to treat cancer, provides a chance to cure many tumors. The perioperative period is characterized by a recovery from cancer status, wherein the side effects and indirect symptoms also affect the physiological status of the body. However, the impact of surgery on the plasma of patients remains unclear. To clarify the changes, we investigated short-term changes in the plasma proteome during the perioperative period. In our cohort, a total of 101 plasma samples were collected during the perioperative period, including first visit (FV, n=17), pre-operative (Pre-op, n=49), and post-operative (Post-op, n=35). Clinical characteristics, surgery information and plasma proteome data are presented in **Table S6.**

A total of 8,394 non-redundant proteins were identified. In the FV, Pre-op and, post-op samples, 6,230, 7,702 and 7,225 proteins were identified in each group, respectively.

To find the difference between pre-op and post-op, we performed differential expressed gene analysis between proteome data of samples before and after operations (t-test) and determined 312 DEPs (p-value <0.05, fold change >2). To further investigate genes with different functions associated with recovery from surgery, we performed the k-means clustering for the DEPs and identified two clusters, of which one has clustered the proteins that are upregulated after the operation, while the other cluster contained the downregulated proteins. Subsequently, Reactome pathway enrichment analysis was performed for each cluster. Cluster 1, consisting of 60 most abundant proteins, showed down-regulation and was mainly involved in the cellular response to heat stress, dissolution of fibrin clots, and DNA methylation, indicating the factors of tumorigenesis were diminished after surgery. Cluster 2 consisted of 252 up-regulated proteins mainly involved in regulating IGF transport and platelet adhesion and, therefore, indicated the stress response to surgery (**Figures 7A, 7B, and 7C**). Importantly, we found the heat shock protein family was significantly altered after surgery, including HSPA1A, HSPA1B, HSPA5, and HSPA9. Among these proteins, HSPA1A ($p = 0.043$), HSPA1B ($p = 0.043$) and HSPA5 ($p = 0.033$) were significantly down-regulated after surgery, while the HSPA9 ($p=0.0028$) was significantly up-regulated (**Figure 7D**). Consistent with the previous reports[37, 38], HSP1A and HSPA9 were identified as unfavorable and favorable prognostic markers, respectively, in colorectal cancer (p-value <0.001).

To discover the markers that indicate the therapeutic effect of the surgery, we focused

on the proteins that were identified in the FV and pre-op but dramatically downregulated in post-op. According to these criteria, 193 significant DEPs were identified (Kruskal-Wallis p-value <0.05, **Figure 7E**). By matching these with the 1,399 TDP specific proteins, 32 were overlapped (**Figure 7F**). These 32 proteins demonstrated the potential to indicate the detection of tumor or the indication of the surgery recovery status (**Figure 7G**). We then accessed the correlation network of these 32 proteins; as shown in **Figure 7H**, the CDK5 has the largest number of significantly correlated proteins, suggesting the involvement of the CDK5 in carcinogenesis. Further, we investigated the protein expression of CDK5 in plasma proteome derived from healthy and tumor patients and found CDK5 was significantly up-regulated in esophageal cancer (p = 1.768e-4), ovarian cancer (p = 1.919e-5), and pancreatic cancer (p = 3.707e-3) (**Figure 7I**).

Since CDK5 is a kinase that plays an important role in regulating the nervous system and cell cycle, we investigated the expression of the substrates proteins of CDK5 and identified that four (APP, DPYSL2, EPRS1, PAK1) of 50 substrates were significantly correlated to CDK5 (p-value < 0.01, **Figures 7J, 7K, S10A, S10B, and S10C**). Moreover, pathway enrichment analysis of the correlated substrates of CDK5 revealed that vascular endothelial growth factor and its receptor (VEGFR2) mediated vascular permeability was significantly enriched (SRC, RAC1, PAK, CTNNB1) (**Figure 7L**). VEGF is a key mediator of tumor angiogenesis and a major target for anti-angiogenic therapy for various malignant tumors. These results suggested tumor-induced angiogenesis was

overexpressed in the tumor, and its diminishment could be monitored after the surgery (**Figure 7M**).

Construction of tumor diagnosis classifier based on proteome data

In this study, using the proteome data of normal plasma as the baseline, we built a reference interval (RI) with the distribution of normal plasma protein abundance, which could be used for abnormal detection if the protein abundance exceeds the upper limit of the RI. For any query sample, an outlier is defined as a protein whose expression level is higher than the upper limit of the RI, and the detection of multiple outliers would further increase the statistical confidence for patients with tumors. To this end, we constructed proteomic RIs derived from 200 normal samples. The upper limit of the RI is defined as P75 + 3*(P75 – P25), where P75 and P25 are the 75^{th} and 25^{th} percentiles of the protein abundance in the normal groups, respectively (**Figure 8A**).

As a proof of concept to illustrate the usage of the RI, we next analyzed the sample pairs from 9 kinds of tumor (n=877) that were identified from 40 patients (LC (n=232), ML (n=138), RCC (n=130), BLCA (n=121), CRCA (n=101), STAD (n=65), TGCT (n=46), and ESCA (n=44)). We mapped the tumor plasma proteomes with the same workflow used in the normal plasma samples (**Figure 8B**). The results showed tumor plasma proteome obtained more outlier than in normal plasma proteome. The average number of outliers in tumor plasma samples were ranged from 247 to 369, while the

number in normal samples ranged from 135 to 155.

Finally, we trained a linear classifier to classify the tumor plasma samples from normal ones using the significance of the outliers. The area under receiver operating characteristic ranged from 0.97 (LM) to 0.99 (CRCA) (**Figure 8C**). Six-fold cross-validation was performed to ascertain the robustness of the classifier. The result showed an average recall of 0.95 of normal samples and 1.0 of tumor samples (**Figure 8C**).

Overall, these findings demonstrate that the landscape of the plasma-derived proteomes can be beneficial in the diagnosis of tumor types and monitoring the therapeutic effect after surgery.

Discussion

Liquid biopsy tests showed promise for early cancer detection, tumor classification, and monitoring treatment response. The plasma in bodily fluids could represent an essential component of the liquid biopsy test. Despite several previous plasma protein biomarker studies, a comprehensive on pan-cancer derived plasma biomarkers was lacking due to limited proteomic datasets and appropriate controls to guide data analysis and interpretation. Here, we performed a large-scale, comprehensive analysis of the human plasma atlas comprising 1,118 human cancer proteome samples obtained from 18 tumor types and 200 healthy human proteome samples. Among the subclusters, various

classical and tumor-specialized subtypes are defined.

We first performed NMF-based unsupervised clustering on all tumor-derived top 3,000 most variable plasma proteins. We emphasized the importance of biological recognition that integrates histological boundaries and unites different original patients. Such concepts can reveal the homo- or heterogeneity among different tumor types and lead to new insights into treatments that are more effective in one group than others. Deeply, quantitative analyses of proteomic data provide a unique landscape with potential clinical effects[39]. For example, we identified new associations between tumor oncogenic signaling and clinical annotations, including upregulation of neutrophil-related clinical annotations and metastasis in the STAD and CRCA of Gastrointestinal-rich cluster, highlighting the link between inflammation and invasion. In this way, the relevant proteins can be identified to validate their biological process (e.g., S100A8 and S100A9). A higher correlation of erythrocyte-related clinical annotations and oxidative condition revealed the intrinsic biological distribution, which provided potential clinical markers in the BLCA of Urinary I-rich cluster. Furthermore, creatinine induced imbalance between metabolic reprogramming and immune environment, the key proteins INSR and GLUT4 demonstrated the dependence of tissues on glucose, which suggested that the therapeutic window for targeting glucose metabolic is broad. These findings demonstrated potential diagnostic or therapeutic directions in the difficult arena of tumor heterogeneity.

We classified fifteen tumor types into six physiological systems, and 1,360 system plasma proteins were identified. Intriguingly, the pathway enrichment analysis revealed that the specific plasma proteins were significantly enriched in the predominant biological processes of corresponding systems. Then, we used hierarchical clustering to divide the 19 tumor types into four subgroups based on their correlations. The results showed that samples from the same physiological system clustered together, indicating more similarity in tissue-originated tumor type. The integrated analysis of the clinical indices associated with proteome subgroups is one of the highlights of this study. In our data, plasma TBA level was the highest in cluster 3, and a significant positive correlation between TBA and interleukin signaling was observed. Interestingly, the proteins APOB, CD36, SFTPA1, JAK1, and SDC1 related to interleukin signaling were up-regulated in cluster 3. Hence, we inferred that the metabolic disorder of TBA might be due to the inflammatory response caused by the abnormal activation of the interleukin pathway in the liver. Besides traditional biomarkers, such as SCCA, CEA, and NSE, we also identified 18 protein modules associated with different tumor types, including 616 potential biomarkers in eight tumor types, of which 116 proteins were exclusively detected in the TDP but never found in the healthy samples. Even more exciting, we identified some proteins associated with the TNM stage, such as CRP, C9, and CFB in the urinary system. These proteins associated with the TNM stage may guide physicians to select appropriate therapeutic schedules. In summary, our study further fortifies the

understanding of cancer biology, revealing new biomarkers and biomarker panels, and opens a new avenue for more efficient early diagnosis and surveillance of cancer.

In recent years, studies of immunotherapy combined with immune checkpoint inhibitors (ICI) have made a great breakthrough, improving the treatment options for various cancers and increasing the survival rate of treated patients, although heterogeneous response rates were obtained for treated patients with different cancer types and patients affected by a specific tumor type[40]. In particular, ICI might also improve the treatment of urothelial cancer, gastric cancer, CRCA, LC, and breast cancer, considering the promising results achieved so far and the relatively low efficacy of currently available treatments[41-44]. By investigating the proteomic profiles of tumors, this study proposes a rational stratification of pan-cancer patients-based immune signatures for personalized therapeutic interventions. Our results supported recent reports that more elevated immune infiltrating tumors have lower lipid metabolism and higher glucose metabolism activities[34]. The *IC5* group exhibited upregulation of platelet activation, platelet plug formation and coagulation cascade and displayed an elevated megakaryocytes level, suggesting megakaryocytes in the TME were villains of venous thrombosis in LC. Moreover, a higher PLT count and FIB associated with *IC5* further supported this finding.

In addition, we identified two clusters of DEPs for the proteins relevant in recovery from cancer status and surgery. The tumor removal caused the reduction of CDK5

signaling and further decreased the VEGFR2 mediated vascular permeability and angiogenesis. The considerable differences in proteome patterns between the Pre-op and Post-op highlighted the probability of referring the protein panel to monitor the therapeutic effect of the surgery. Proteins that display increasing or decreasing intensities are more likely to be of interest for future studies.

We constructed the prediction model to classify the tumor patients and normal control based on the plasma proteome. We built the reference interval of normal people, trained the linear classifier to predict the tumor types based on the plasma proteome. Taken together, our findings revealed that the plasma proteome could reflect the diverse molecular mechanisms of different tumor types. The plasma proteins could be used as biomarkers for early-stage cancer detection and treatment response.

Conclusions

In summary, this study extends our biological understanding of plasma-derived proteins of pan-cancer and generates therapeutic hypotheses that may serve as the basis for future preclinical studies and clinical trials toward molecularly guided precision treatment of pan cancer. Meanwhile, we have made the primary and processed datasets available in publicly accessible data repositories and portals, which will allow full investigation of this extensively characterized cohort by broader scientific communities.

Abbreviations

TCGA: The Cancer Genome Atlas

MSKCC: Memorial Sloan Kettering Cancer Center

CPTAC: Clinical Proteomic Tumor Analysis Consortium

CEA: carcinoembryonic antigen

MS: mass spectrometry

DIA: data-independent acquisition

LC: lung cancer

ML: malignant lymphoma

RCC: renal carcinoma

BLCA: bladder cancer

BRCA: breast cancer

CRCA: colorectal cancer

STAD: stomach adenocarcinoma

TGCT: testicular germ cell tumors

ESCA: esophageal carcinoma

OVCA: ovarian cancer

CECA: cervical carcinoma

LIHC: liver hepatocellular carcinoma

CHOL: cholangiocarcinoma

SARC: sarcoma

PAAD: pancreatic adenocarcinoma

THCA: thyroid carcinoma

HNCA: head and neck carcinoma

TDP: tumor-derived plasma

NTDP: non-tumor-derived plasma

PLT: platelet

RBC: red blood cell

HGB: hemoglobin

FOT: fraction of total

FDR: false discovery rate

iBAQ: intensity-based absolute quantification

NMF: non-negative matrix factorization

CV: coefficients of variation

TTES: tumor type enrichment score

CCP: ConsensusCLusterPlus

ORA: over-representation analysis

DEPs: differentially expressed proteins

HPMP: hydrogen peroxide metabolic process

ssGSEA: single-sample gene set variation analysis

GSVA: Gene Set Variation Analysis

APR: acute phase response

NE%: neutrophil percentage

MGPS: Multi-Gene Proliferation Scores

CREA: creatinine

LY: lymphocyte count

KEGG: Kyoto Encyclopedia of Genes and Genomes

TBA: total bile acid

ALB: albumin

BTB: blood-testis barrier

WGCNA: Weight Gene Correlation Network Analysis

VTE: thromboembolism

FIB: fibrinogen

FV: first visit

Pre-op: pre-operative

Post-op: post-oprative

GO: gene ontology

FFPE: formalin-fixed paraffin-embedded

RI: reference interval

Methods

Patient samples of Pan-Cancer

Clinical sample acquisition

The pan-cancer plasma samples used in this study were obtained from the Affiliated Hospital of Hebei University. Plasma samples were collected from patients or healthy controls. Clinical information of 1,128 patients, including tumor type, gender, age, tumor node metastasis (TNM) staging, and biochemical indicators, is listed in Table S1. All human samples included in the present study were obtained after approval of the Research Ethics Committees of the Affiliated Hospital of Hebei University (HDFY- LL-2021-050), and the institutional review board of Department of Urology of Fudan University Shanghai Cancer Center (2005-ZZK-25), together with the written informed consent from each patient and healthy controls. Patients were excluded if they have been treated with radiotherapy or chemotherapy or suffering from other cancers.

Proteomic Workflow

Plasma protein extraction and trypsin digestion

The top 14 highest abundant plasma proteins were first depleted from plasma samples before protein extraction using an immunodepleting kit (Thermo Fisher) according to the

manufacturer's instructions and then inactivated at 85°C for 10 min. The depleted plasma was digested by trypsin at an enzyme to protein mass ratio of 1:25 overnight at 37°C, and the peptides were then extracted and dried (SpeedVac, Eppendorf).

FFPE Protein extraction and tryptic digestion

Slides (10 μm thick) from FFPE blocks were dissected according to Haemotoxylin and Eosin (H&E) staining, deparaffinized with xylene and washed with gradient ethanol. Approximately 1 mg human tumor sample (wet weight tissue) was homogenized in 200 μL lysis buffer [0.1 M Tris-HCL (pH 8.0), 0.1 M DTT, and 1 mM PMSF]. The samples were ground by grinding rods for 3 min, and 50 μL 20% SDS was added to reach a maximum SDS concentration of 4%. Lysates were boiled at 99°C, 1,500 rpm for 0.5, 1, and 1.5 h and then centrifuged at 12,000 × g for 5 min at 25°C. Lysate supernatant was transferred into acetone at a maximum ratio of 1:3 and kept at −20 °C for at least 4 h. Precipitated proteins were washed with cooled acetone three times and redissolved in 8 M urea. Before digestion, samples were loaded into the FASP tubes. Digestion was performed with FASP tube with trypsin in 50 μL ammonium bicarbonate (ABC, 50 mM) for 16 h at a 1:50 enzyme-to-protein ratio at 37°C. Digested samples were collected by centrifugation at 12,000 × g for 15 min and washed twice with 200 μL of mass spectrometry water. Samples were desalted on C18 columns and dried down using centrifugation at 12,000 × g.

Nano-LC-MS/MS

Samples were measured using LC-MS instrumentation consisting of an EASY‐nLC

1200 ultra-high-pressure system (Thermo Fisher Scientific) coupled via a nano-electrospray ion source (Thermo Fisher Scientific) to a Fusion Lumos Orbitrap (Thermo Fisher Scientific). The peptides were dissolved with 12 µL loading buffer (0.1% formic acid in water), and 5 µL was loaded onto a 100 µm I.D. × 2.5 cm, C18 trap column at a maximum pressure 280 bar with 14 µL solvent A (0.1% formic acid in water). Peptides were separated on 150 µm I.D. × 15 cm column (C18, 1.9µm, 120 Å, Dr. Maisch GmbH) with a linear 15–30% Mobile Phase B (ACN and 0.1% formic acid) at 600 nL/min for 75 min. The MS analysis was performed following a data-independent approach (DIA). The DIA method consisted of an MS1 scan from 300–1,400 m/z at 60k resolution (AGC target 4e5 or 50 ms). Then, 30 DIA segments were acquired at 15k resolution with an AGC target of 5e4 or 22 ms for maximal injection time. The setting "inject ions for all available parallelizable time" was enabled. High energy collision dissociation (HCD) fragmentation was set to a normalized collision energy of 30%. The spectra were recorded in profile mode. The default charge state for the MS2 was set to 3.

Peptide identification and protein quantification

All data were processed using Firmiana (Feng et al., 2017). DIA data were searched against UniProt human protein database (updated on 2019.12.17, 20406 entries) using FragPipe (v12.1) with MSFragger (2.2) (Kong et al., 2017). The mass tolerances were 20 ppm for precursor and 50 mmu for product ions. Up to two missed cleavages were allowed. The search engine was set with cysteine carbamidomethylation as a fixed modification and N-acetylation and oxidation of methionine as variable modifications. Precursor ion score charges were limited to +2, +3, and +4. The data were also searched

against a decoy database so that protein identifications were accepted at a false discovery rate (FDR) of 1%. The results of DIA data were combined into spectra libraries using SpectraST software. A total of 327 libraries were used as reference spectra libraries.

DIA data was analyzed using DIA-NN (v1.7.0) (Demichev et al., 2020). The default settings were used for DIA-NN (Precursor FDR: 5%, Log lev: 1, Mass accuracy: 20 ppm, MS1 accuracy: 10 ppm, Scan window: 30, Implicit protein group: genes, Quantification strategy: robust LC (high accuracy)). Quantification of identified peptides was calculated as the average of chromatographic fragment ion peak areas across all reference spectra libraries. Label-free protein quantifications were calculated using a label-free, intensity-based absolute quantification (iBAQ) approach (Zhang et al., 2012). We calculated the peak area values as parts of corresponding proteins. The fraction of total (FOT) was used to represent the normalized abundance of a particular protein across samples. FOT was defined as a protein's iBAQ divided by the total iBAQ of all identified proteins within a sample. The FOT values were multiplied by 10^5 for ease of presentation, and missing values were imputed with 10^{-5}.

Proteome Data Preprocess

Mass spectrometry platform QC

The 293T cell (National Infrastructure Cell Line Resource) lysate was measured every three days as the quality-control standard for the quality control of the performance of mass spectrometry. The quality-control standard was digested and analyzed using the same method and conditions as the tumor FFPE samples. A pairwise Spearman's correlation coefficient was calculated for all quality-control runs in the statistical analysis

environment R v3.6.3. The results are shown in Figure S1. The average correlation coefficient among the standards was 0.91, and the minimum and maximum correlation coefficients were 0.84 and 0.97, respectively. These quality-control samples demonstrated the consistent stability of the mass spectrometry platform.

Preprocessing of DIA proteomic data and batch correction

Before performing any downstream analysis, we applied type-specific batch correction on global proteome abundance to remove the technical difference. A batch correction was performed for each data type on the subset of proteins with more than 20% observed in at least one subtype. After filtering proteins with missing rates of >20% in each tumor type, there were 5286 proteins with an overall missing rate of 40.9% to 61.5% for the diagnosis-wide global protein abundance datasets, respectively. As the first step to batch correction, we performed KNN imputation separately on the data from each tumor type using the "impute.knn" function from the "impute" R package. After merging the data across tumor type, we then applied the R tool Combat, with the tumor type as a covariate to remove batch effects[45].

Quantification and statistical analysis

Unsupervised proteome clustering using NMF

We used non-negative matrix factorization (NMF) implemented in the NMF R-package[46] to perform unsupervised clustering of tumor samples and identify proteome features showing characteristic expression patterns for each cluster. Briefly, given a factorization

rank k (where k is the number of clusters), NMF decomposes a p x n data matrix V into two matrices W and H such that multiplication of W and H approximates V. Matrix H is a k x n matrix whose entries represent weights for each sample (1 to N) to contribute to each cluster (1 to k), whereas matrix W is a p x k matrix representing weights for each feature (1 to p) to contribute to each cluster (1 to k). Matrix H was used to assign samples to clusters by choosing the k with the maximum score in each column of H. The resulting matrix was then subjected to NMF analysis leveraging the NMF R-package. To determine the optimal factorization rank k (number of clusters) for the proteome data matrix, we tested a range of clusters between k = 2 and 10. For each k, we factorized matrix V using 50 iterations with random initializations of W and H. To determine the optimal factorization rank, we calculated cophenetic correlation coefficients measuring how well the intrinsic structure of the data was recapitulated after clustering and chose the k with maximal cophenetic correlation for cluster numbers between k = 2 and 10. After determining the optimal factorization rank k, we repeated the NMF analysis using 200 iterations for clustering.

The tumor type distribution of NMF clusters

To determine the distribution of different tumor types in each NMF clusters, we calculated the tumor type enrichment score (TTES) using the following formula:

$$TTES_{ij} = \frac{T_{ij}/Tj}{T_i/T_a}$$

$TTES_{ij}$ represents the TTES of specific tumor type i in the j NMF cluster. Specifically, T_{ij}, Tj, T_i, T_a represents the count of tumor type i in the j NMF cluster samples, the count of j NMF cluster samples, the count of i tumor type samples, and all tumor samples, respectively.

Pathway analysis based on NMF clusters

To better determine the intrinsic proteomic clusters and pathway activation, unsupervised clustering was performed. Based on NMF cluster proteome data, features (proteins) were filtered according to the Kruskal-Wallis test, finally, 3326 proteins with Benjamin and Hochberg adjusted p-value<0.05 across 1118 samples were selected for clustering. Consensus clustering was performed using the R package ConsensusClusterPlus [47]. Before clustering, the data matrix was scaled so that each protein had a mean 0 and standard variation 1 across samples and calculated the mean value of each protein based on the NMF cluster. Then, K-means clustering based on the Euclidean distance matrix was conducted across 300 repetitions for cluster numbers ranging from 6 to 20, and other parameters were used as default. To determine the optimal protein groups, we performed an overrepresentation of biological pathway/gene sets in each proteome group using the clusterProfiler R package[48]. The final protein groups (k = 9) were chosen to maximize the number of significant pathway associations based on the Hallmark gene sets from MSigDB. Based on the nine protein groups, a further comprehensive analysis was

performed based on Reactome[49], KEGG[50], Hallmark[51] and GOBP[52] gene sets.

Pathway Consolidation via Sumer

wing to the significant redundancy of gene membership across the number of gene sets, the gene set enrichment results are hard to interpret. To reduce the complexity of the redundancy results, we utilized the Sumer tool[53]. This tool uses an affinity propagation algorithm to cluster similar pathway gene sets into largely distinct modules. Sumer was run based on the minus log10 transform of q-value, derived from the clusterProfiler results as weights. Consolidated pathway modules for each gene protein were identified based on the top 50 pathways by weight (Table S2).

Gene set Score for Single Sample

To functional characterized NMF cluster results by single sample Gene Set Enrichment Analysis (ssGSEA), we calculated the normalized enrichment score of each sample based on four class of gene sets: GOBP, KEGG, Hallmark and Reactome gene sets. We utilized R package GSVA with following parameters: min.sz = 10, max.sz = 300 and other parameters were used default.

Multi-Gene Proliferation Scores (MGPS)

MGPS were calculated from the median-MAD normalized proteome data. Briefly,

MGPS was calculated as the mean expression level of all cell cycle-regulated genes identified by Whitfield et al. [54]in each sample.

Immune Scores

The immune score was inferred using the R package ESTIMATE v1.0.11 [33]. Although the ESTIMATE algorithm was designed to analyze transcriptome data, some studies have used it for proteome analysis[55, 56]. The results indicate the feasibility of evaluating the engagement of each subtype of immune cells.

Immune subtype identification

The abundance of 64 different cell types was computed via xCell based on proteomic profiles. Therefore, for this analysis, 1118 tumor plasma samples with proteome data were utilized. Consensus clustering was performed based on Raw enrichment scores of the patients. Based on these 64 signatures, consensus clustering was performed to identify groups of samples with similar immune/stromal characteristics. Consensus clustering was performed using the R packages ConsensusClusterPlus based on z-score normalized signatures. Specifically, 80% of the original pan-cancer tumor plasma samples were randomly subsampled without replacement and partitioned into three major clusters using the Partitioning Around Medoids (PAM) algorithm, which was repeated 2000 times.

Correlation between tumor types and clinical features

To estimate the correlations between tumor types and clinical features, Fisher's exact test was used on categorical variables, and Wilcoxon rank-sum test was used on continuous

variables.

Hierarchical clustering

The Hclust function implemented in the R language was used to perform unsupervised clustering of Pan-cancer samples to identify proteomic features of each cluster. Before clustering, we selected proteins as follows: 1) Within each tumor type, proteins were required to express in at least 1/2 samples. 2) Within each tumor type, the filtered proteins were sorted in descending order by mad (Median Absolute Deviation). 3) The 20% bottom mad proteins of each tumor type were combined, and duplicates were removed, resulting in 5,286 proteins. These 5,286 proteins were averaged in each tumor type, resulting in a 5,286 x 19 protein-expression matrix sorted in descending order by mad. Four clear clusters were found when using the Pearson algorithm based on the 1,124 top mad proteins.

Screening potential drug targets of pan-cancer samples

Drug targets approved by the FDA or under clinical trials were retrieved from the Drugbank database (version 5.1.5) (http://www.drugbank.ca/). Target proteins that were unregulated in pan-cancer with potential curative drugs were chosen.

Classifier construction

The classifier was implemented using Scikit Learn (1.4.1). Normalization was performed by the scale function. Reference interval (RI) was built with the distribution of protein abundance of normal samples. The upper limit of the RI was defined as P75 + 3 × (P75

– P25), where P75 and P25 are the 75th and 25th percentile of protein abundance in the normal samples, respectively. Next, Fisher's exact test was used to calculate the p-value of each sample. The linear classifier was used to classify tumor patients to normal samples using the significance of drug response outliers. Data visualization was implemented in matplotlib (3.1.1).

Declarations

Ethical Approval and Consent to participate

All human samples included in the present study were obtained after approval of the Research Ethics Committees of the Affiliated Hospital of Hebei University (HDFY- LL-2021-050), and the institutional review board of Department of Urology of Fudan University Shanghai Cancer Center (2005-ZZK-25), together with the written informed consent from each patient and healthy controls.

Consent for publication

All authors give consent for the publication of the manuscript in Molecular Cancer.

Availability of data and materials

Data resources: The accession number for the MS proteomics data reported in this paper is IPX0003227001. The link access to the raw data as follows:

https://www.iprox.cn/page/SSV024.html;url=1626408275641N9P8

The password: juUc

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key R&D Program of China (2017YFA0505102 to C. D., 2016YFA0502500 to C. D., 2018YFA0507501 to Z. Q, and 2017YFC0908404 to Z. Q.), the Outstanding young scientific research and innovation team of Hebei University (605020521007), the National Natural Science Foundation of China (31770886 to C. D., 31972933 to C. D., and 31700682 to Z. Q.), the Science and Technology Commission of Shanghai Municipality (2017SHZDZX01 to C. D.), and the Major Project of Special Development Funds of Zhangjiang National Independent Innovation Demonstration Zone (ZJ2019-ZD-004 to C. D.), Postdoctoral Science Foundation(Grant No. 2020T130114 to J. F.), the 65th China Postdoctoral Science Foundation to Y. W., the Fudan original research personalized support project to Z. Q,.

Authors' contributions

Study Conception & Design, C.D., Y.J., D.Y., L.Y.; Project Administration, L.B., C.D., Y.J., D.Y., L.Y.; Sample Resources, X.Q., Y.Q., Y.Z., H.G., L.L., A.Z., W.D., Y.W., L.W., X.W., Y.L., J.L., X.Y., G.Z., D.L., X.G.; Patient Sample Management and QC, X.Q., Y.Q., Y.Z., H.G.; Experiments & Data Collection, L.Z., J.F., L.B., J.L.; Data Analysis, L.B., J.F., J.L., G.Y.; Writing, L.B., J.F., J.L., G.Y., C.D.

Acknowledgements

We thank members of DC Lab, for collecting the clinical information.

Authors' information

Department of Urology, Fudan University Shanghai Cancer Center, State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Science, Institute of Biomedical Sciences, and Human Phenome Institute, Fudan University, Shanghai 200433, China;

Lin Bai, Jinwen Feng, Yuanyuan Qu, Jiacheng Lyu, Guojian Yang, Lingli Zhu, Dingwei Ye, Chen Ding

Department of Medical Oncology, Affiliated Hospital of Hebei University; Hebei Key Laboratory of Cancer Radiotherapy and Chemotherapy, 212 Yuhua East Road, Baoding, Hebei 071000, P.R. China;

Yuanxue Zhu, Aimin Zang, Xiaofang Wang, Yumiao Li, Guofa Zhao, Youchao Ji

Department of Oncology, Shanghai Medical College, Shanghai 200032, China;

Yuanyuan Qu, Dingwei Ye

College of Pharmaceutical Sciences, Key Laboratory of Medicinal Chemistry and Molecular Diagnosis, Ministry of Education, Hebei University, Baoding 071002, China;

Xiaoqiang Qiao

Hebei Normal University of Science & Technology, Qinhuangdao, 066004, China;

Hui Gao

Department of Neurosurgery, The Affiliated Nanhua Hospital, University of South China, Hengyang, 421002, China;

Lingxiao Liao

Clinical Lab, Affiliated Hospital of Hebei University, Baoding, Hebei 071000, China;

Yan Wang

Department of Hematology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China;

Liang Wang

Department of hepatological surgery, Affiliated Hospital of Hebei University, 212 Yuhua East Road, Baoding, Hebei 07100, P.R. China;

Jinghua Li

CT/MRI room, Affiliated Hospital of Hebei University, 212 Yuhua East Road, Baoding, Hebei 071000, P.R. China;

Xiaoping Yin

College of Clinical Medicine, Hebei University, Baoding 071002, China; Department of Medical Oncology, Affiliated Hospital of Hebei University, Baoding 071000, China;

Dan Liu, Xiangpeng Gao

Figure Legends

Figure 1. Technological aspects of plasma protein profiling.

A. Workflow of 1,118 Pan-cancer collection, processing, sequencing, and data analysis. **B**. Patient-centric circos plot representing the key demographic and histologic features of 1,118 Pan-cancer patients. Grey gaps represent missing data. Numbers to the right indicate samples in each of the categories. **C**. Number of proteins quantified in each tissue. Each dot represents data from one tumor type. **D**. The Venn diagram shows the overlap of proteins identified in tumor-derived plasma (TDP) and non-tumor-derived plasma cases (NTDP). **E.** Distribution of the number of proteins quantified across different numbers of tumors. **F**. Dynamic ranges and total of $\log_{10}$(FOT) in 18 tumor types **G**. Principal component analysis (PCA) plot of 18 TDP and NTDP proteomes (indicated by colors) obtained from 1,118 Pan-cancer and 200 NTDP cases. **H**. Different types of proteins were detected between tumor and normal samples. **I**. The upper Venn diagram shows the overlap of drug-related proteins identified in tumor and normal samples. The lower Venn diagram shows the overlap of transcription factors (TFs) identified in tumor and normal samples.

See also Figure S1and Table S1.

Figure 2. Proteomic classification revealed heterogeneity of pan-cancer.

A. Proteomic clusters and differentially expressed proteins assigned to nine gene groups

(top heatmap). Each row represents a proteomic cluster, and each column represents a protein. Red/blue represent up/downregulation patterns of different proteins in a cluster. Tumor type enrichment score (top-right heatmap), distributions of tumor types, clinical outcomes, and genders among the six clusters (top-left pie plots) and gene members of key pathways enriched in each gene group (bottom heatmap) are shown. For each pathway, the averaged single-sample gene set enrichment analysis (ssGSEA) score in each proteomic cluster based on global proteomics (protein) data are paneled on the bottom-right. **B**. Distribution of TNM stages among six clusters. **C**. Distribution of TNM stages between STAD, CRCA and other tumor types. **D.** Stacked bar plot showing the fraction of tumor node metastasis (TNM) stages of STAD and CRCA which belong to Gastrointestinal-rich cluster and other clusters, separately. **E.** GSEA (GOBP gene sets) analysis revealed Acute phase response (APR) was enriched in STAD and CRCA which belong to Gastrointestinal-rich cluster than other clusters (FDR<0.25). **F.** Heatmap of APR-related proteins (FDR<0.05) in STAD and CRCA between Gastrointestinal-rich cluster and other clusters, values were standardized by z-score. **G.** Boxplot showing the distribution of CRP (Ranksums p-value<0.05) in STAD and CRCA between Gastrointestinal-rich cluster and other clusters. **H.** The spearman correlation of CRP and clinical annotations in STAD and CRCA. **I.** Boxplot showing the distribution of neutrophil percentage (NE%) (Ranksums p-value<0.05) in STAD and CRCA between Gastrointestinal-rich cluster and other clusters. **J.** The spearman correlation of NE%

validated neutrophils-related metastasis proteins in STAD and CRCA. **K and L**. Boxplot showing the distribution of S100A8 and S100A9 (Ranksums p-value<0.05) in STAD and CRCA between Gastrointestinal-rich cluster and other clusters. **M**. Diagram showing the proteome profiles of APR, neutrophil and its effect on metastasis.

See also Figure S2 and Table S2.

Figure 3. The character of the Urinary I-rich cluster.

A. The levels of hemoglobin level (HGB) in clinical annotation were different among six non-negative matrix factorization (NMF) clusters; p-value was calculated by the Kruskal-Wallis test. **B**. Boxplot depicting the distribution of HGB in clinical annotation between BLCA and other tumors in Urinary I-rich, the p-value was derived by Ranksums test. **C**. Same layout as (B), showing the distribution between Urinary I-rich and other clusters in BLCA. **D**. Heatmap illustrating the gender fraction between Urinary I-rich and other clusters in BLCA. **E.** Same layout as (B), showing the distribution among different groups: Urinary I-rich Male, Other clusters Male, Other clusters Female. **F**. GSEA analysis (Hallmark and GOBP gene sets) depicting the enrichment in heme metabolism and hydrogen peroxide metabolic process (HPMP) in male BLCA in Urinary I-rich than in other clusters (FDR<0.25) **G.** Heatmap illustrating the proteins participate in heme metabolism and HPMP in male BLCA between Urinary I-rich and other clusters, values were standard by z-score. **H**. Regression scatterplot showing the high spearman

correlation between HGB level in clinical annotation and heme metabolism. **I.** Same layout as H, showing the correlation between heme metabolism and HPMP. **J**. The correlation between HPMP and Multi-Gene Proliferation Scores (MGPS). **K**. Boxplot depicting the distribution of Multi-Gene Proliferation Scores (MGPS) between Urinary I-rich and other clusters in male BLCA. **L.** Diagram of HGB effects in male BLCA of Urinary I-rich. **M.** The levels of CREA in clinical annotations were different among six NMF clusters, p-value was based on Kruskal-Wallis test. **N**. Scattering plot illustrating the higher correlation of CREA and pancreas β cells. **O**. Regression scatterplot showing the high spearman correlation between pancreas β cells and response to insulin. **P**. Same layout as (M), showing the correlation between insulin receptor signaling pathway and glucose metabolic process, glucose import. **Q**. Heatmap illustrating the proteins participate in glucose metabolic process and insulin receptor signaling pathway in Urinary II-rich and other clusters, values were standard by z-score. **R.** Diagram of CREA affects glucose metabolism by activating insulin related-pathway in Urinary II-rich.
See also Figure S3 and Table S2.

Figure 4. Plasma proteome-based hierarchical clustering of pan-cancer.

A. Clustering of enriched proteins in the six physiological systems to identify proteins that can specify the biological systems. **B.** KEGG function enrichment analysis of the system-specific plasma proteins. **C.** The identification of specific system plasma proteins

associated with TNM staging (Kruskal-Wallis test, $p<0.05$). **D** Heat map showing the pairwise correlation between all 18 tumor types based on plasma protein expression levels. The average protein values for each protein and tumor type are used in the analysis. Four subtype clusters (C1, C2, C3, C4) detected based on the correlation between the 18 tumor types. **E** KEGG analysis of all 1,124 differentially expressed proteins (DEPs) of the four subtype clusters. **F** Comparison of plasma ALB level in the four clusters. Kruskal-Wallis test, P-value = 2.803E-7. **G** Comparison of plasma ALB level among the five tumor types (TGCT, HNCA, THCA, PAAD, and CHOL). Kruskal-Wallis test, P-value<0.0001. **H** Pathways correlated with ALB level. Spearman's correlation, P-value<0.05. **I.** Comparison of pathway scores for the cell junction assembly among the five tumor types. Kruskal-Wallis test, P-value = 1.7E-07. **J.** The correlations of the Cell Junction Assembly and Rac Protein Signal Transduction. Spearman's correlation, R = 0.24, P-value<0.05. **K.** Proteins correlated with Rac Protein Signal Transduction. Proteins were distinguished by the pathways they were involved in. Spearman's correlation, P-value<0.05. **L.** Heat map showing the high level in TGCT than other tumor types in Cluster 4. **M.** RAC signaling might hyper-stimulate the BTB formation by increasing the activation of cell junction assembly, which prevents the ALB transport in TGCT.

See also Figure S4 and Table S3

Figure 5. Identification of specific tumor-derived plasma proteins.

A. Seven conventional plasma protein markers identified in different tissue types. **B.** Density plots showing the expression of CA125 and CEA in different tumor types. **C.** KEGG analysis of the tumor-specific plasma proteins. **D.** Module-trait relationship. Separation of the proteomic profiles to eigengene modules identified a group of modules positively associated with different tumor types. P-value<0.05, R>0.92. The right panel showing the protein number of each module. **E.** The top 10 functional hub proteins in the modules with the highest connectivity are shown and annotated with potential clinical utilities. **F.** The identification of specific RCC plasma potential biomarkers positively associated with the TNM stage. Kruskal-Wallis test, p<0.05. **G.** The identification of specific RCC plasma potential biomarkers negatively associated with the TNM stage. Kruskal-Wallis test, p<0.05. **H.** The bar plot showing the proteins with potential biomarker value and biological relevance in tissue proteome overlapped with plasma proteome. **I.** The Venn diagram shows the overlap of potential biomarker proteins identified in lung tissue and lung plasma-derived. **J.** KEGG analysis of the potential specific plasma markers in lung cancer (LC). **K.** Exclusively expressed proteins in LC-derived plasma.

See also Figures S5 and S6 and Table S4.

Figure 6. Immune infiltration of 1,118 Pan-cancer.

A. Heatmap illustrating cell type compositions and activities of selected pathways across five immune clusters. The heatmap in the first section illustrates the immune signatures. ssGSEA scores based on global proteomics data for biological pathways upregulated in different immune groups are illustrated in the remaining sections. **B**. Rose Charts showed different tumor types among the five immune clusters. **C**. Distribution of the immune score of five immune subtypes. **D.** The correlation between pathway score of lipids metabolism and immune score. **E**. Heatmap showing the comparison between immune clusters (rows) with NMF clusters (columns). **F.** The correlation between pathway score of glucose metabolism and immune score. **G**. Heatmap showing proteins of lipids metabolism and glucose metabolism among five immune clusters. **H**. The influence of immune infiltration by lipid and glucose metabolism. **I.** Distribution of megakaryocytes from xCell in the five immune subtypes. **J.** The correlation between megakaryocytes from xCell and proteins. **K.** Heatmap showing proteins of platelet activation, crosslinking collagen fibrils, and platelet plug formation among the five immune clusters. **L.** Distribution of platelet counts (PLT) in five immune subtypes. **M.** The correlation between platelet counts (PLT) and fibrinogen (FIB). **N**. Distribution of FIB in five immune subtypes. **O**. A model depicting the megakaryocytes level was associated with diverse clinical features, may reveal the molecular mechanism influenced on thrombus production in LC plasma.

See also Figure S7 and Table S5.

Figure 7: Short-term changes of the plasma proteome after surgery.

A. Heatmap of k-means clustering of the DEPs during the perioperative period. **B.** Functional annotation of genes in two clusters. **C**. Boxplot of representative DEPs in plasma before and after surgery. **D**. Heatmap of expression of HSPA in plasma during the perioperative period. **E**. Volcano plot of DEPs in plasma before and after surgery. **F.** The Venn diagram of down-regulated proteins after surgery versus tumor-derived plasma-specific proteins. **G**. Heatmap of proteins expression of 32 common proteins of the down-regulated proteins after surgery and tumor-derived plasma-specific proteins. **H.** Network of co-expression network of 32 common proteins. Node size indicates the number of proteins correlated to the node. Edges indicate the interaction between proteins. **I.** Boxplot of CDK5 expression in plasma before versus after surgery, ESCA versus normal, OVCA versus normal, PAAD versus normal. **J**. Correlation and significance between CDK5 and substrates of CDK5. **K**. Scatter plot of pairwise change of protein expression of CDK5 and PAK1 before and after surgery. **L**. Heatmap of expression of proteins in CDK5 correlated angiogenesis pathway. **M**. A proposed pathway of CDK5 angiogenesis pathway.

See also Figure S8 and Table S6.

Figure 8. Proteomic classifier to predict tumor patients.

A. Schematic diagram of building a reference interval for proteins of normal samples. **B.** Schematic diagram of using reference intervals to predict tumor samples. **C.** The area under the receiver operating characteristic (AUROC) of the predictor of 9 kinds of tumors (LC, ML, CRCA, RCC, TGCT, STAD, ESCA, BRCA, and BLCA).

Figure S1. Sample collection criteria and quality assessments for proteome data, related to Figure 1.

A. Sample collection process. **B**. Pie charts showing the proportion of 18 tumor types and non-tumor-derived plasma (NTDP). **C**. A spearman's correlation coefficient was calculated for 50 quality control (QC) runs using HEK293T cell samples (range, 0.82-0.99). **D**. Box plots of proteins identified in tumor types and NTDP; p-value was calculated by an unpaired t-test. **E**. The Venn diagram shows the overlap of different types of proteins identified in tumor and normal samples. **F**. Barplots showing protein counts of different types of proteins among 18 tumor types and NTDP.

Figure S2. Molecular subtyping and the characterization of gastrointestinal tumors, related to Figure 2.

A. Cophenetic correlation coefficient (y-axis) calculated for a range of factorization ranks (x-axis). The maximal and optimal cophenetic correlation coefficient was observed for rank K = 6, shown in red. **B.** Silhouette plot for K = 6. This plot indicates the quality of cluster separation. **C**. Line plot showing the number of MSigDB C2 enriched gene sets

among different gene groups which clustered by CCP. The maximal point was observed for nine gene groups, shown in red. **D**. The panel of stacked bar plot showing the distribution of some clinical annotations (Gender and TNM stages) among tumor types in each NMF cluster. **E**. Boxplots depicting the distribution of proteins validated as neutrophils-related metastasis between STAD and CTC, which belong to Gastrointestinal-rich cluster and other clusters, separately. p-values were derived by Ranksums tests.

Figure S3. Characterization of urinary tumor, related to Figure 3.

A. The distribution of erythrocyte-related clinical annotations among six NMF clusters, p-value were calculated by Kruskal-Wallis tests. **B**. Boxplot revealing the distribution of HGB between LIHC and other tumor types in Urinary I-rich cluster, combining with Figure 2A and Figure 3B, we conclude that BLCA, not LIHC, is characterized by a high level of HGB in Urinary I-rich cluster. **C.** The comparison of fraction in T stage (Early-stage: T0 and T1, Late-stages: T2, T3, T4) between Urinary I-rich and other clusters in BLCA. **D**. Kaplan-Meier curves showing survival outcome of male BLCA patients with a high level of HBA2 in TCGA cohort. The p-value was derived from the log-rank test. **E.** Same layout as (D), showing the female survival outcomes. **F.** The distribution of insulin receptor signaling pathway among six NMF clusters, p-value were calculated by Kruskal-Wallis test. **G.** Same layout as (F), showing the level of LYM# from clinical

annotation. **H**. Regression scatterplot illustrating the spearman correlation between glucose metabolic process and LYM#. **I.** Heatmap depicting the higher level of pathways: pancreas β cells, glucose import, and glucose metabolic process in Urinary II-rich cluster.

Figure S4. Specific-system plasma proteins associated with TNM staging, related to Figure 4.

A–C. The identification of specific-system plasma proteins associated with TNM staging in excretory system, reproductive system, respiratory system, respectively (Kruskal-Wallis test, $p < 0.05$).

Figure S5. Characterization of LIHC, related to Figure 4.

A. Comparison of plasma TBA level in the four clusters. Kruskal-Wallis test, P-value = 2.2E-6. **B.** Comparison of plasma TBA level in LIHC and SARC. Kruskal-Wallis test, P-value = 0.0009. **C.** Proteins correlated with TBA. Proteins were distinguished by the pathways they were involved in. Spearman's correlation, $p < 0.05$. **D.** The heat map shows the expression level of FDPS and CYP27A1 in different tumor types. **E.** The expression of FDPS in the TGCA LIHC cohort. **F.** The activated lipid metabolism leads to TBA biosynthesis disorder by upregulating the FDPS expression level. **G** The expression of S100A10, RAB29 and CNN2 in the TGCA cohort. **H.** Comparison of pathway scores for the branched-chain amino acid metabolism among the five tumor

types (TGCT, CHOL, HNCA, PAAD, and THCA). Kruskal-Wallis test, P-value = 1.7E-06.

Figure S6. Identification of specific tumor-derived tissue proteins, related to Figure 5.

A. Heatmap for non-ubiquitous (3,107), ubiquitous (2,179) proteins. The color bar on the right indicates the relative expression abundance. **C.** The workflow for the protein extraction process of tumor FFPE samples.

Figure S7 and S8.The potential biomarker in tumor types, related to Figure 5.

The upper Venn diagram shows the overlap of potential biomarker proteins identified in tissue and plasma-derived samples. KEGG analysis of the potential specific plasma markers in lung cancer. The lower heatmap showed exclusively expressed proteins never found in healthy control but found in cancer-derived plasma.

Figure S9. Immune infiltrations of 1,118 pan-cancer, related to Figure 6.

A. Consensus matrices of the 1,118 tumor samples from k = 2 to k = 5. **B.** Cumulative distribution function plot, delta plot, and corresponding tracking plot. **C**. Distribution of metabolism of lipids among five immune clusters.

Figure S10. Pairwise foldchange of CDK5 and substrates of CDK5 after surgery.

A. Scatter foldchange of CDK5 and positively correlated substrates of CDK5 after surgery. **B.** Scatter plot of foldchange of CDK5 and negatively correlated substrates of CDK5 after surgery. **C.** Scatter plot of the fold change of CDK5 and all substrates of CDK5 after surgery.

Reference

1. Aravanis, A.M., Lee, M. & Klausner, R.D. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574 (2017).
2. Hanash, S.M., Ostrin, E.J. & Fahrmann, J.F. Blood based biomarkers beyond genomics for lung cancer screening. *Transl Lung Cancer Res* **7**, 327-335 (2018).
3. Petralia, F. et al. Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell* **183**, 1962-1985.e1931 (2020).
4. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341.e1323 (2017).
5. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine* **23**, 703-713 (2017).
6. Kandoth, C. et al. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73 (2013).
7. Xu, J.Y. et al. Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **182**, 245-261.e217 (2020).
8. Jiang, Y. et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **567**, 257-261 (2019).
9. Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science translational medicine* **6**, 224ra224 (2014).
10. Thierry, A.R., El Messaoudi, S., Gahan, P.B., Anker, P. & Stroun, M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev* **35**, 347-376 (2016).
11. Kilgour, E., Rothwell, D.G., Brady, G. & Dive, C. Liquid Biopsy-Based Biomarkers of Treatment

Response and Resistance. *Cancer cell* **37**, 485-495 (2020).
12. Bronkhorst, A.J., Ungerer, V. & Holdenrieder, S. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomolecular detection and quantification* **17**, 100087 (2019).
13. Barbie, D.A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112 (2009).
14. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
15. Yang, L. et al. DNA of neutrophil extracellular traps promotes cancer metastasis via CCDC25. *Nature* **583**, 133-138 (2020).
16. Jaillon, S. et al. Neutrophil diversity and plasticity in tumour progression and therapy. *Nature reviews. Cancer* **20**, 485-503 (2020).
17. Coffelt, S.B., Wellenstein, M.D. & de Visser, K.E. Neutrophils in cancer: neutral no more. *Nature reviews. Cancer* **16**, 431-446 (2016).
18. Coffelt, S.B. et al. IL-17-producing γδ T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* **522**, 345-348 (2015).
19. Fiorito, V., Chiabrando, D., Petrillo, S., Bertino, F. & Tolosano, E. The Multifaceted Role of Heme in Cancer. *Front Oncol* **9**, 1540 (2019).
20. Sies, H. & Jones, D.P. Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. *Nature reviews. Molecular cell biology* **21**, 363-383 (2020).
21. Solis, M.Y., Artioli, G.G. & Gualano, B. Potential of Creatine in Glucose Management and Diabetes. *Nutrients* **13** (2021).
22. Tang, N. et al. Correlation analysis between four serum biomarkers of liver fibrosis and liver function in infants with cholestasis. *Biomedical reports* **5**, 107-112 (2016).
23. Liu, Z. et al. Prognostic Value of the CRP/Alb Ratio, a Novel Inflammation-Based Score in Pancreatic Cancer. *Ann Surg Oncol* **24**, 561-568 (2017).
24. Ijichi, C., Matsumura, T., Tsuji, T. & Eto, Y. Branched-chain amino acids promote albumin synthesis in rat primary hepatocytes through the mTOR signal transduction system. *Biochem Biophys Res Commun* **303**, 59-64 (2003).
25. Bart, J. et al. An oncological view on the blood-testis barrier. *The Lancet. Oncology* **3**, 357-363 (2002).
26. Mack, N.A. et al. β2-syntrophin and Par-3 promote an apicobasal Rac activity gradient at cell-cell junctions by differentially regulating Tiam1 activity. *Nature cell biology* **14**, 1169-1180 (2012).
27. Felder, M. et al. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol Cancer* **13**, 129 (2014).
28. Wettersten, H.I., Aboud, O.A., Lara, P.N., Jr. & Weiss, R.H. Metabolic reprogramming in clear cell renal cell carcinoma. *Nature reviews. Nephrology* **13**, 410-419 (2017).
29. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
30. Yano, Y. et al. Expression and localization of ecto-nucleotide pyrophosphatase/phosphodiesterase

I-3 (E-NPP3/CD203c/PD-I beta/B10/gp130RB13-6) in human colon carcinoma. *International journal of molecular medicine* **12**, 763-766 (2003).

31. Yano, Y. et al. Expression and localization of ecto-nucleotide pyrophosphatase/phosphodiesterase I-1 (E-NPP1/PC-1) and -3 (E-NPP3/CD203c/PD-Ibeta/B10/gp130(RB13-6)) in inflammatory and neoplastic bile duct diseases. *Cancer Lett* **207**, 139-147 (2004).
32. Aran, D., Hu, Z. & Butte, A.J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**, 220 (2017).
33. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612 (2013).
34. Hubler, M.J. & Kennedy, A.J. Role of lipids in the metabolism and activation of immune cells. *The Journal of nutritional biochemistry* **34**, 1-7 (2016).
35. Vitale, C. et al. Venous thromboembolism and lung cancer: a review. *Multidisciplinary respiratory medicine* **10**, 28 (2015).
36. Kattula, S., Byrnes, J.R. & Wolberg, A.S. Fibrinogen and Fibrin in Hemostasis and Thrombosis. *Arteriosclerosis, thrombosis, and vascular biology* **37**, e13-e21 (2017).
37. Jiang, W., Pan, X., Yan, H. & Wang, G. Prognostic Significance of the Hsp70 Gene Family in Colorectal Cancer. *Med Sci Monit* **27**, e928352 (2021).
38. Lee, S.L. et al. in Heat Shock Protein-Based Therapies. (eds. A.A.A. Asea, N.N. Almasoud, S. Krishnan & P. Kaur) 345-379 (Springer International Publishing, Cham; 2015).
39. Anderson, L. Six decades searching for meaning in the proteome. *Journal of proteomics* **107**, 24-30 (2014).
40. Sharma, P. et al. The Next Decade of Immune Checkpoint Therapy. *Cancer Discov* **11**, 838-857 (2021).
41. Kono, K., Nakajima, S. & Mimura, K. Current status of immune checkpoint inhibitors for gastric cancer. *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association* **23**, 565-578 (2020).
42. Almquist, D.R., Ahn, D.H. & Bekaii-Saab, T.S. The Role of Immune Checkpoint Inhibitors in Colorectal Adenocarcinoma. *BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy* **34**, 349-362 (2020).
43. Lin, A., Wei, T., Meng, H., Luo, P. & Zhang, J. Role of the dynamic tumor microenvironment in controversies regarding immune checkpoint inhibitors for the treatment of non-small cell lung cancer (NSCLC) with EGFR mutations. *Mol Cancer* **18**, 139 (2019).
44. Swoboda, A. & Nanda, R. Immune Checkpoint Blockade for Breast Cancer. *Cancer treatment and research* **173**, 155-165 (2018).
45. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* **8**, 118-127 (2007).
46. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
47. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).

48. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**, 284-287 (2012).
49. Fabregat, A. et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-d655 (2018).
50. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-d361 (2017).
51. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
52. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
53. Savage, S.R., Shi, Z., Liao, Y. & Zhang, B. Graph Algorithms for Condensing and Consolidating Gene Set Analysis Results. *Molecular & cellular proteomics : MCP* **18**, S141-s152 (2019).
54. Whitfield, M.L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977-2000 (2002).
55. Clark, D.J. et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964-983.e931 (2019).
56. Zhang, H. et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765 (2016).