Molecular Therapy
Oncolytics
Original Article

Integrating SWATH-MS Proteomics and Transcriptome Analysis Identifies CHI3L1 as a Plasma Biomarker for Early Gastric Cancer

Li Min,[1] Shengtao Zhu,[1] Rui Wei,[1] Yu Zhao,[1] Si Liu,[1] Peng Li,[1] and Shutian Zhang[1]

[1]Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease, Beijing Digestive Disease Center, Beijing Key Laboratory for Precancerous Lesion of Digestive Disease, Beijing 100050, P. R. China

Early diagnosis of gastric cancer (GC) provides patients opportunities for minimally invasive endoscopic resection. Here, we developed a new strategy integrated the state-of-the-art sequential windowed acquisition of all theoretical fragment ion (SWATH) mass spectra (MS) with multi-dataset joint analysis to screen for the stage-I GC plasma biomarker. In SWATH-MS assays, we identified 37 upregulated and 21 downregulated proteins in GC plasma. In the mRNA database analysis, 633 genes were identified as differentially expressed genes in at least 4 out of 5 datasets, but there were only 94 genes identified as upregulated. Only 1 gene, CHI3L1, was characterized as upregulated in both the dataset consensus list and the SWATH-MS list. Then, we detected the CHI3L1 level in the plasma of a large cohort consisting of 200 participants. The area under the ROC curve (AUC) of CHI3L1 in distinguishing GC from others was 0.788. Integrating the plasma CHI3L1 level with clinical factors further boosted the AUC to 0.887. In conclusion, we provide a novel strategy for biomarker screening, combining recent MS techniques with public database analysis, and identified plasma CHI3L1 as a potential biomarker for patients with endoscopically resectable GC.

INTRODUCTION

With 1,033,701 new cases and 782,685 deaths worldwide, gastric cancer (GC) ranks second in mortality and sixth in incidence among all cancer types in 2018.[1] Incidence rates of GC are markedly elevated in regions in East Asia as compared to North America and Europe.[2] Noticeably, GC exhibits a specific high incidence in China, with an estimated 679,100 new cases in 2015, and about 498,000 Chinese patients died from GC in the same year.[3] Many studies suggested that the prognosis of GC has a very close association with the stage at detection.[4] Overall survival of GC dramatically decreased if the diagnosis did not occur at an early stage.[5] The 5-year survival rate of GC is quite low in China, because more than 80% of GC patients were diagnosed at an advanced stage.[6] Additionally, early diagnosis of GC could not only improve the prognosis but also avoid the trauma of open surgery and improve the quality of life for stage-T1a patients who were suitable for endoscopic, minimally invasive treatment.[7,8] Thus, an increase in the early detection rate of GC is urgently needed. Recently, gastric endoscopy is the main approach to screening for GC, which showed a very promising effect in improving GC patients' mortality.[9] However, according to the recent guidelines, most gastric endoscopy screenings were conducted in people roughly selected by non-specific risk factors such as age, family history, and *Helicobacter pylori* infection.[10–13] Inefficient pre-selection of high-risk individuals resulted in low cost efficiency and low compliance, which largely weakened the application of endoscopy screening in developing countries. Thus, the discovery of new biomarkers to identify high-risk individuals is of vital importance. During the past decade, many efforts have been made in GC biomarker discovery, mainly focused on plasma and serum.[14] The concept of liquid biopsy is now commonly accepted, along with the wide application of high-throughput technologies and microfluidic technologies in biomarker discovery.[15] Most attention was given to circulating cell-free DNA (cfDNA). However, cfDNA was supposed to be released into plasma during tumor cell death, which was suitable for the monitoring of recurrence rather than early diagnosis.[16] The plasma levels of some proteins, such as MG7Ag, S100A9, GIF, and AAT, also have been associated with the diagnosis or prognosis of GC.[14,17–20] However, most of those biomarkers were discovered using a cohort of mixed stages, and none of them were validated in endoscopically curable T1a GC patients.

Sequential windowed acquisition of all theoretical fragment ion (SWATH) mass spectra (MS) is emerging as a new strategy of data-independent acquisition (DIA) MS methods, which allows a complete

Received 19 February 2020; accepted 26 March 2020;
https://doi.org/10.1016/j.omto.2020.03.020.

Correspondence: Shutian Zhang, PhD, Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease, Beijing Digestive Disease Center, Beijing Key Laboratory for Precancerous Lesion of Digestive Disease. No. 95, Yong'an Road, Xicheng District, Beijing 100050, P.R. China.
E-mail: zhangshutian@ccmu.edu.cn
Correspondence: Peng Li, MD, PhD, Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease, Beijing Digestive Disease Center, Beijing Key Laboratory for Precancerous Lesion of Digestive Disease. No. 95, Yong'an Road, Xicheng District, Beijing 100050, P.R. China.
E-mail: lipeng@ccmu.edu.cn

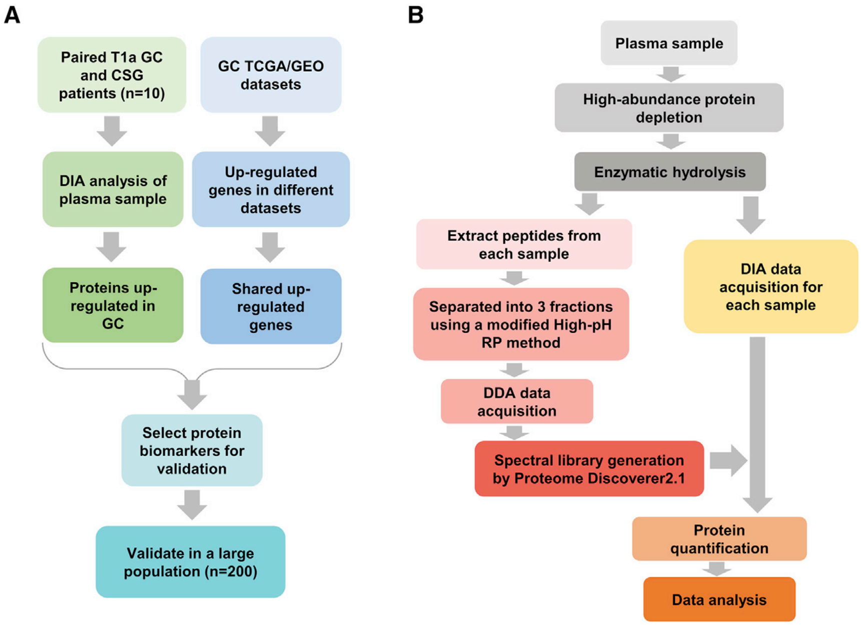
A
Paired T1a GC and CSG patients (n=10)
GC TCGA/GEO datasets
DIA analysis of plasma sample
Up-regulated genes in different datasets
Proteins up-regulated in GC
Shared up-regulated genes
Select protein biomarkers for validation
Validate in a large population (n=200)
B
Plasma sample
High-abundance protein depletion
Enzymatic hydrolysis
Extract peptides from each sample
Separated into 3 fractions using a modified High-pH RP method
DDA data acquisition
Spectral library generation by Proteome Discoverer2.1
DIA data acquisition for each sample
Protein quantification
Data analysis

Figure 1. General Analysis Pipeline

(A) Flow chart for study design. (B) Workflow of SWATH-MS analysis.

and permanent recording of all fragments of detectable peptides in each sample.[21] SWATH-MS combines deep proteome coverage capabilities with quantitative accuracy and consistency and becomes the state-of-the-art MS platform in cancer biomarker discovery.[22] Recently, SWATH-MS was already applied in screening biomarkers of prostate carcinoma, esophageal carcinoma, and hepatocellular carcinoma.[23–25] After an accumulation of a large amount of cancer sequencing data in the past 2 decades, many gene expression databases, such as the Cancer Genome Atlas (TCGA)[26] and Gene Expression Omnibus (GEO),[27] have become more and more popular in cancer biomarker discovery studies. Those publicly accessible repertoires not only allow us to re-analyze and reinterpret previously published data but also give us independent cohorts for verification of our own findings.

Chitinase 3-Like 1 (CHI3L1) is a carbohydrate-binding lectin with an affinity for chitin but does not have chitinase activity. CHI3L1 is associated with tissue remodeling, inflammation, and dendritic cell accumulation.[28,29] CHI3L1 could be secreted by many cell types—including activated macrophages, fibroblasts, chondrocytes, neutrophils, vascular smooth muscle cells, and synovial cells—and was reported associated with fibrosis, asthma, and many different types of cancer.[29–32] It was reported that CHI3L1 protein functions as an oncogenic protein in breast cancer,[33] colorectal cancer,[34] and GC.[35] However, its potential function as a diagnositic biomarker of GC has not yet been revealed.

In this study, we used SWATH-MS proteomics to screen for specific upregulated proteins in plasma of stage-I GC patients and then filtered the candidates by public GC gene expression datasets and the human secretome list (Figure 1).[36] With very strict screening criteria, only CHI3L1 was identified as a robust biomarker candidate, which was further verified in an independent cohort with a large population. Our study provided a novel strategy in cancer biomarker discovery combining a high-resolution MS technique with publicly available mRNA profiling database analysis, and identified plasma CHI3L1 as a robust biomarker for patients with endoscopically resectable GC.

RESULTS

Identification of Aberrantly Elevated Proteins in GC Plasma by SWATH-MS

To identify specific plasma protein biomarkers in GC patients, we subjected 10 plasma samples from paired T1a GC and chronic superficial gastritis (CSG) patients to highly abundant protein depletion (Figure S1A) and SWATH-MS analysis. For data-dependent acquisition (DDA) and spectral library generation, those specimens from all 10 subjects were pooled together and divided into 3 fractions using a modified high-pH reversed-phase (RP) method, and DDA was performed (Figures S1B and S1C). A total of 9,070 peptides, which mapped to 1,170 proteins, were identified (Figures S2A–S2D).

For DIA and protein quantification (Figures S3 and S4), protein abundance in each sample was well distributed with a mean mass error of 1.0 ppm, suggesting that very robust results were acquired in DIA (Figure S5A). 873 proteins were quantified in all 5 GC samples (Figure 2A), and 877 proteins were quantified in all 5 CSG samples (Figure 2B). No protein was characterized as GC or CSG specific (Figure S5B). With a threshold fold change (FC) $\geq$ 1.2 or $\leq$ 0.83 combined with a false discovery rate (FDR) $\leq$ 0.1, we identified 37 proteins upregulated and 21 proteins downregulated in GC plasma (Figure 2C; Table S1). The differentially abundant proteins could well distinguish GC plasma from CSG plasma (Figure 2D).

Differentially Abundant Plasma Proteins in GC Are Mutually Connected

Noticeably, 10 out of 58 differentially abundant proteins could not be mapped to a typical protein-coding gene, and most of them were fragments of circulating antibodies. Then, we annotated the remaining 48 up- and downregulated protein-coding genes using Gene Ontology (GO) terms (Figure 3A). For GO: BP (biological process), nearly half of those genes were associated with response to stress and the immune system process. For GO: CC (cell component), there were 25 genes associated with extracellular space and 20 genes associated with the plasma membrane. For GO: MF (molecular function), 24 genes were associated with ion

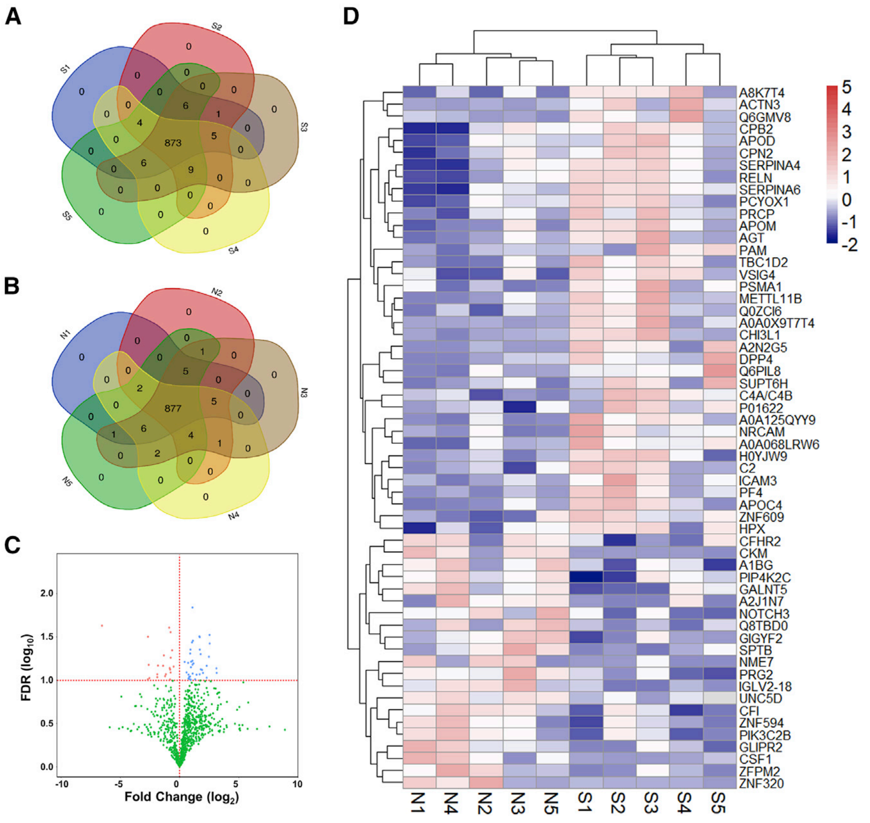

Figure 2. Identification of GC Plasma-Specific Proteins by SWATH-MS

(A) SWATH-MS identified and quantified 873 proteins in all 5 GC samples. (B) SWATH-MS identified and quantified 877 proteins in all 5 CSG samples. (C) Scatterplot exhibited aberrantly expressed proteins in GC plasma. Blue points, upregulated; red points, downregulated; green points, non-significant. (D) The differentially abundant proteins could well distinguish GC plasma from CSG plasma by hierarchical clustering.

binding. ClueGO function analysis was performed to give insight into the biological process in which differentially expressed genes (DEGs) were involved. At the medium level, regulation of humoral immune response, positive regulation of macrophage-derived foam cell differentiation, and regulation of extracellular matrix organization were identified (Figure 3B; Table S2). Detailed-level ClueGO function analysis was also conducted, and the results are shown in Figure S6 and Table S3. Additionally, the protein-protein interaction (PPI) relationship was also revealed by the STRING database. Except for 5 scattered genes, most of those genes were associated with each other (Figure 3C). Among all those genes, Angiotensinogen (AGT), Hemopexin (HPX), CHI3L1, Serpin Family A Member 4 (SERPINA4), and so forth were identified as hub genes (number of connections > 5).

The Consensus of Upregulated Proteins in GC based on Different Online Datasets

Given the assumption that the tumor tissue could be the most possible source of proteins elevated in GC plasma,[36] here we performed DEG analysis in five different GC datasets using GEO2R to construct a consensus list of upregulated proteins in GC for further filtering the biomarker candidates identified by plasma SWATH-MS.

There were 14708, 8365, 6942, 1409, and 1304 DEGs in TCGA and GEO: GSE54129, GSE37023, GSE29272, and GSE26942, respectively (Figure 4A). We chose the genes exhibited in at least 4 different DEG lists (663 genes in total) for GO annotation (Figure 4B) and found that they were mostly associated with the cellular process (GO: BP), cell part (GO: CC), and protein binding (GO: MF). Among all those 633 genes, only 10 genes were upregulated in all 5 databases (Figure 4C), suggesting that those GC databases had a high heterogeneity due to different populations and the adoption of different mRNA profiling techniques. ClueGO function analyses at global (Figure S7), medium (Figure 5A), and detailed (Figure S8) levels were all performed. Cell differentiation, adhesion, migration, response to oxygen, extracellular structure organization, and so forth were found associated with those DEGs (Figure 5A; Tables S4 and S5). Then, we performed PPI analysis in all genes upregulated in at least 3 databases (94 genes) by STRING. Most of those genes were also associated with each other, except for some small gene clusters and scattered genes (Figure 5B).

Identification and Verification of CHI3L1 as an Elevated Biomarker in GC Plasma

Considering that most specific plasma protein biomarkers were either secreted proteins or membrane proteins, we filtered the potential biomarkers identified by plasma SWATH-MS with the database-derived upregulated gene list and a widely recognized secreted/membrane protein list.[36] 58 out of 94 database-derived upregulated genes were secretome or membrane associated (Figure 6A), and only 1 gene, CHI3L1, was characterized as upregulated in both the database-derived list and SWATH-MS (Figures 6B and 6C).

To verify the potential diagnostic effect of CHI3L1 in early GC patients, the plasma of 100 GC patients, 25 benign neoplasia patients (BN), 43 patients with hyperplastic polyps (HPS), and 32 normal controls (NC) was subjected to CHI3L1 enzyme-linked immunosorbent assay (ELISA) (Table 1). We found that the concentrations of CHI3L1 in plasma were 162.59 ± 112.78, 87.41 ± 50.33, 64.50 ± 35.44, and 76.22 ± 29.49 in patients with GC, patients with BN, patients with HP, and NC, respectively (Figure 6D). Thus, CHI3L1

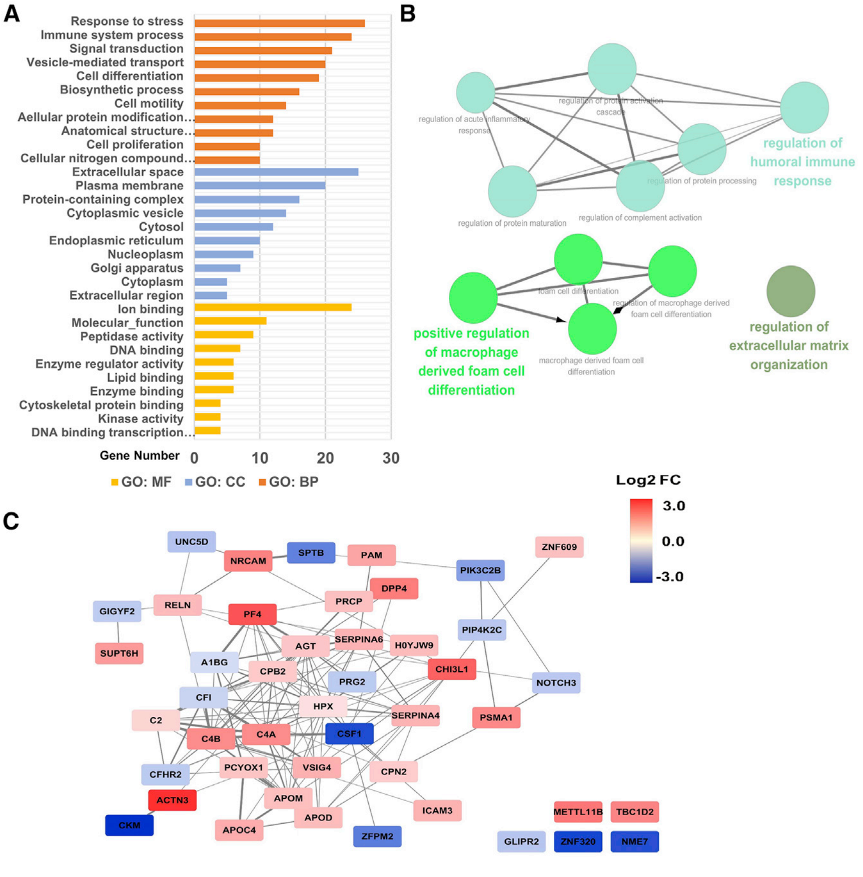

Figure 3. Functional Analysis of Differentially Abundant Plasma Proteins in GC

(A) Statistics of GO terms annotated to differentially abundant plasma proteins in GC. BP category is indicated in red, CC category is indicated in blue, and MF category is indicated in yellow. (B) GO: BP terms related to regulation of humoral immune response, positive regulation of macrophage-derived foam cell differentiation, and regulation of extracellular matrix organization were enriched by ClueGO function analysis at medium level. (C) A protein-protein interaction (PPI) network among the differentially abundant plasma proteins was constructed by the STRING database, and the color of the nodes (proteins) was mapped to their $\log_2$ GC/CSG fold changes.

was significantly higher in the plasma of GC patients, as compared to BN patients ($t = 4.974$, $df = 123$, $p < 0.001$), patients with HPs ($t = 7.843$, $df = 141$, $p < 0.001$), and NCs ($t = 6.952$, $df = 130$, $p < 0.001$). Additionally, we found that CHI3L1 was positively associated with age (Figures S9A and S9B), but not sex (Figures S9C and S9D), in both GC patients and healthy individuals. We also found that the GC patients of different stages showed a similar plasma CHI3L1 level (Figure S9E).

Identification of CHI3L1 as a Robust Plasma Biomarker in GC Patients

Diagnostic accuracy of the plasma CHI3L1 for predicting GC was assessed by receiver operating characteristic (ROC) curve analysis, and the area under the ROC curve (AUC) value was calculated. We demonstrated that CHI3L1 could distinguish GC patients from those with precancerous BN, patients with HPS, and NC, with AUCs of 0.722, 0.844, and 0.764, respectively. The overall AUC comparing GC to all other participants was 0.788 (Figure 6E). Then we stratified GC patients with the clinical stage. The plasma CHI3L1 level showed considerable performance in identifying stage-I GC patients, with AUCs of 0.708 (versus that for BN), 0.831 (versus that for HPS), 0.744 (versus that for NC), and 0.773 (versus that for all others), respectively (Figure S10A). For the stage-II/III patients, the AUCs slightly increased to 0.773 (versus that for BN), 0.893 (versus that for HPS), 0.841 (versus that for NC), and 0.846 (versus that for all others) (Figure S10B).

The association between CHI3L1 level and clinical factors such as age and sex could bias the prediction of GC by CHI3L1 level. Thus, we used a logistic model to integrate plasma CHI3L1 level with other clinical factors to achieve a better diagnostic effect in identifying GC. The AUC further increased into 0.887 when integrating plasma CHI3L1 level with age and sex (Figure S11A) in identifying GC patients from all other participants. The efficiencies in distinguishing GC from patients with BN, patients with HPS, and NC were all highly increased when adding age and sex into the prediction model (Figures S11B–S11D).

Then we focused on patients with early-stage, endoscopically resectable GC. Our logistic model integrating plasma CHI3L1 level, age, and sex exhibited an excellent potential as an early GC biomarker, with an AUC of 0.862, in identifying patients with endoscopically resectable stage-I GC from all other participants (Figure 7A). The AUC values of this model were 0.795, 0.902, and 0.862 in distinguishing patients with endoscopically resectable stage-I GC from those with BN, those with HPS, and NC, respectively (Figures 7B–7D). Hence, we suggested that plasma CHI3L1 protein level could serve as a promising predictive biomarker for patients with early-stage, endoscopically resectable GC.

DISCUSSION

Advance in high-throughput technologies such as next-generation sequencing (NGS) and high-resolution mass spectrometry largely promoted our understanding of cancer biology and the development of novel cancer biomarkers and drug targets. During the past decade, large amounts of gene profiling data have been produced and deposited on the website. Effectively exploiting and integrating these publicly accessible big data with state-of-the-art omics techniques could greatly improve the robustness and repeatability of the results. Here, we developed a strategy integrating SWATH-MS proteomics

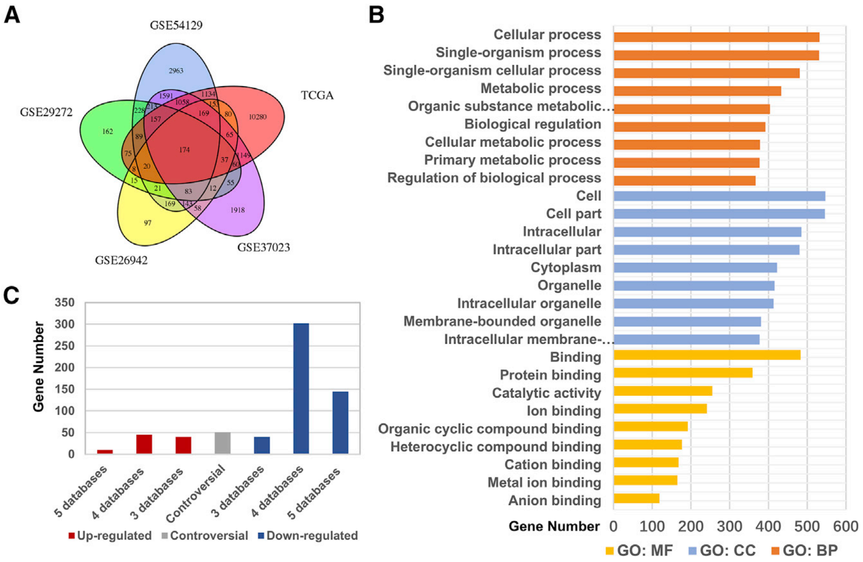

Figure 4. The Consensus of GC Upregulated Proteins by Analysis of Different Datasets

(A) A Venn diagram showed that there were only 174 DEGs shared among all 5 datasets, and 633 DEGs shared among at least 4 datasets. (B) Statistics of GO terms annotated to differentially expressed proteins identified by analysis of GC datasets. BP category is indicated in red, CC category is indicated in blue, and MF category is indicated in yellow. (C) Among all those 633 genes, only 95 genes were characterized as upregulated.

profiling of circulating proteins with a consensus DEG repertoire supported by five different mRNA profiling datasets to find a robust protein biomarker for liquid biopsy.

Unlike other biomarker studies that included GC patients of mixed stages, we further focused on patients with endoscopically curable stage-I GC. All GC patients enrolled in SWATH-MS screening were stage-I GC, and most (79/100) GC patients used in ELISA verification were stage I. As far as we know, it is the largest endoscopically curable GC cohort used in biomarker study. Considering that more than 80% of GC patients were diagnosed at an advanced stage in developing areas, it is hard to recruit a larger early-stage GC population for biomarker discovery.

Using the up-to-date SWATH-MS DIA screening for upregulated proteins in GC plasma, we identified 37 upregulated and 21 downregulated proteins. Those genes were associated with regulation of humoral immune response, positive regulation of macrophage-derived foam cell differentiation, and regulation of extracellular matrix organization, which is in accordance with the related biological processes reported in GC tumor tissues.[37] Among all the upregulated proteins identified in plasma DIA analysis, 7 of them are neither secreted protein nor membrane-associated protein. Most of those proteins (5/7) were nucleic proteins (METTL11B, PSMA1, SUPT6H, TBC1D2, and ZNF609), which could be derived from the apoptotic bodies released by the apoptotic cells to the plasma. Notably, besides CHI3L1, other hub genes such as AGT, HPX, and SERPINA4, identified in the analysis of SWATH-MS results, were all reported associated with cancer in previous publications.[38–40]

In our DIA analysis, all proteins found in the GC subgroup were also detected in at least 1 CSG sample and vice versa. Additionally, there were no sample-specific proteins detected. As shown in Figures 1A and 1B, within both groups, most proteins (873/904, 877/904) were detected in all 5 samples, and every protein species was detected in at least 3 samples. The possible explanation could be the high sensitivity of DIA technology, which could detect proteins in very low abundance.

In the re-analysis of public mRNA databases, there were 14708, 8365, 6942, 1409, and 1304 DEGs identified in TCGA and GEO: GSE54129, GSE37023, GSE29272, and GSE26942, respectively. Only 174 DEGs were shared among all datasets, showed an unsatisfying consistency. The inconsistency of DEGs generated by different databases would be due to the high heterogeneity of different GC populations.[26] To avoid a large amount of false-negative judgments caused by overly stringent criteria, we chose the genes exhibited in at least 4 different DEG lists (663 genes in total) for further analysis. Interestingly, more than 75% of those consensus DEGs were downregulated in GC patients, which further increased the difficulty in the discovery of GC biomarkers.

Here, we identified only 1 secreted protein, CHI3L1, upregulated in both the database consensus list and the SWATH-MS list. CHI3L1 is a 40-kDa glycoprotein that is reported upregulated in some cancer types[29,33–35] as well as some non-neoplastic disease. Previously, CHI3L1 was reported as a biomarker to predict the severity of osteoarthritis, rheumatoid arthritis (RA), inflammatory bowel disease (IBD), and liver fibrosis, most of which were associated with chronic inflammation and fibrosis.[28,41] For cancer-related studies, CHI3L1 was considered as an unfavorable prognostic factor in many cancers, such as breast cancer, colon cancer, lung cancer, and so forth.[29,30,32,33,41] However, no studies revealed the possible application of CHI3L1 as an early diagnostic biomarker. Here, we first proposed that plasma CHI3L1 could be a potential diagnostic marker for early GC. Considering the generality of CHI3L1's biological activity, we could not determine whether CHI3L1 is a pan-cancer biomarker or a GC-specific biomarker. However, individuals with aberrantly elevated plasma CHI3L1 should be advised for further endoscopic screening of GC.

The AUC of GC versus other non-cancerous populations varies from 0.722 to 0.844, suggesting that CHI3L1 is a promising biomarker in identifying GC. The efficiencies in the stage-II/III population (0.773–0.893) are slightly better than those in the stage-I population (0.708–0.831), which is in accordance with the fact that CHI3L1 is prognostic

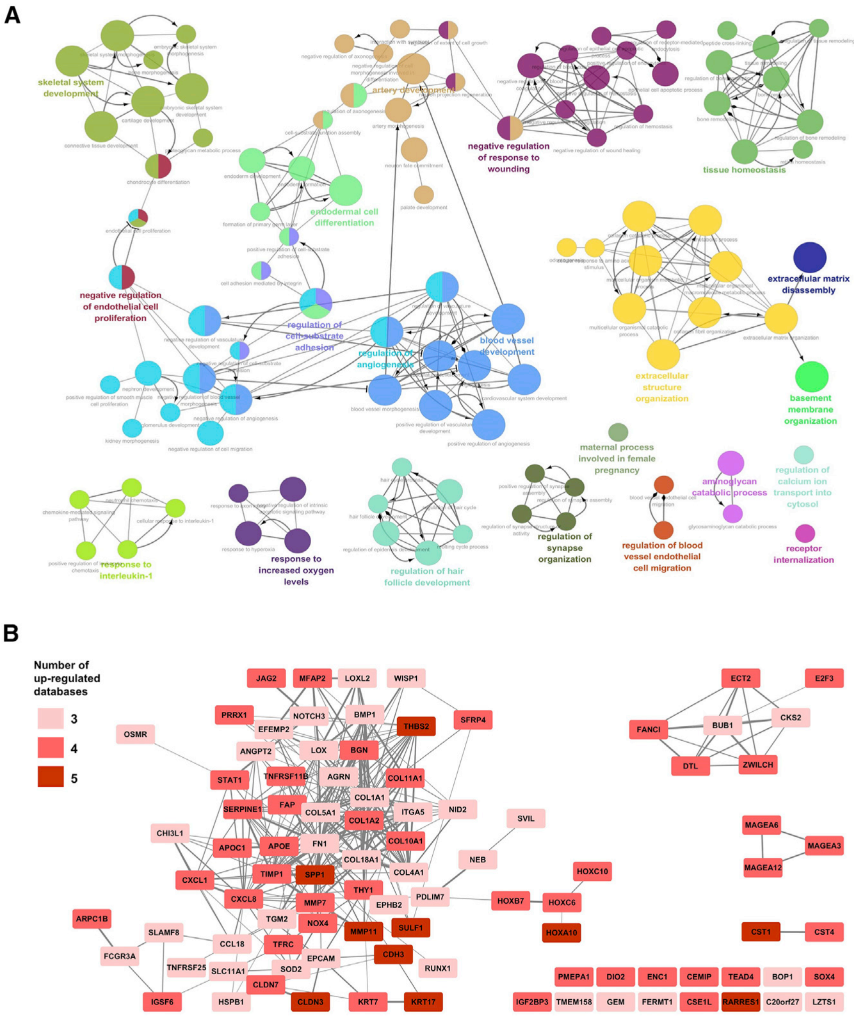
A
skeletal system development
artery development
negative regulation of response to wounding
tissue homeostasis
endodermal cell differentiation
negative regulation of endothelial cell proliferation
regulation of cell-substrate adhesion
regulation of angiogenesis
blood vessel development
extracellular matrix disassembly
extracellular structure organization
basement membrane organization
maternal process involved in female pregnancy
aminoglycan catabolic process
regulation of calcium ion transport into cytosol
response to interleukin-1
response to increased oxygen levels
regulation of hair follicle development
regulation of synapse organization
regulation of blood vessel endothelial cell migration
receptor internalization
B
Number of up-regulated databases
3
4
5
JAG2
MFAP2
LOXL2
WISP1
PRRX1
NOTCH3
BMP1
THBS2
SFRP4
OSMR
EFEMP2
ANGPT2
LOX
BGN
STAT1
TNFRSF11B
AGRN
COL11A1
COL1A1
SERPINE1
FAP
COL5A1
COL1A2
ITGA5
NID2
SVIL
CHI3L1
APOC1
APOE
FN1
COL10A1
COL18A1
NEB
CXCL1
TIMP1
SPP1
COL4A1
THY1
PDLIM7
CXCL8
MMP7
EPHB2
ARPC1B
TGM2
NOX4
MMP11
SULF1
SLAMF8
CCL18
TFRC
CDH3
FCGR3A
EPCAM
TNFRSF25
SLC11A1
SOD2
RUNX1
CLDN7
IGSF6
HSPB1
CLDN3
KRT7
KRT17
ECT2
E2F3
FANCI
BUB1
CKS2
DTL
ZWILCH
MAGEA6
MAGEA3
MAGEA12
HOXC10
HOXB7
HOXC6
HOXA10
CST1
CST4
PMEPA1
DIO2
ENC1
CEMIP
TEAD4
BOP1
SOX4
IGF2BP3
TMEM158
GEM
FERMT1
CSE1L
RARRES1
C20orf27
LZTS1

Figure 5. Functional Analysis of Upregulated Proteins Identified by GC Datasets

(A) GO: BP terms related to cell differentiation, adhesion, migration, response to oxygen, extracellular structure organization, etc., were enriched by ClueGO function analysis at medium level. (B) A protein-protein interaction (PPI) network among the upregulated proteins identified by dataset analysis was constructed by the STRING database, and the color of the nodes (proteins) was mapped to their $\log_2$ GC/NC fold changes.

in many cancers.[29,30,32,33,41] Additionally, when age and sex—two easy-to-access risk factors—were integrated, the AUCs were boosted to 0.862 (versus that for all others), 0.795 (versus that for BN), 0.902 (versus that for HPS), and 0.862 (versus that for NC). The performance of our model in identifying stage-I GC was comparable to that of other models for mixed-stage GC in previous publications,[14,17,18] suggesting that this model is very promising in early GC screening.

CHI3L1 could be secreted by cancer cells, but it could also be secreted by activated macrophages, fibroblasts, and so forth.[28] Thus, we could not conclude whether the elevated plasma CHI3L1 was derived from tumor tissue or generated accompanied with the systemic responses of the human body against tumor. CHI3L1 was known to bind both proteins and carbohydrates, which facilitated potential interactions with the cell surface and extracellular matrix.[41] However, no specific cell-surface-binding partner for CHI3L1 has been identified. Understanding the structure and biochemistry characteristics of CHI3L1 would achieve a specific detection of CHI3L1 with a low cost and further promote the application of CHI3L1 in clinical scenarios in the future.

In conclusion, we developed a novel strategy for disease biomarker screening combining plasma MS technique with public gene expression database analysis and identified plasma CHI3L1 level as a potential biomarker for the identification of patients with endoscopically resectable early-stage GC. Additionally, we proposed a logistic model combining plasma CHI3L1 level with age and sex information, which showed an AUC of 0.862 in distinguishing early-stage GC patients from NC. Our study not only provided a highly sensitive, minimally invasive approach for the identification of patients with endoscopically resectable GC but also gave a paradigm of integrating publicly accessible gene expression datasets with custom-made omics screening procedures in the discovery of disease biomarkers.

MATERIALS AND METHODS

Patient Information and Sample Collection

Five stage-IA (T1aN0M0) GC patients who received endoscopic resection of GC at the Department of Gastroenterology, Beijing Friendship Hospital, Beijing, China, between February 2017 and March 2017 were enrolled in this study along with 5 age- and sex-matched non-cancerous CSG volunteers as controls for biomarker discovery. Additionally, 100 GC patients, 25 BN patients, and 43 patients with HPs received endoscopic or surgical resection of gastric lesions at the Department of Gastroenterology, Beijing Friendship Hospital between March 2017 and December 2018, and 32 healthy volunteers were enrolled in this study for biomarker evaluation. All clinical data are summarized in Table 1. Blood specimens were collected from each patient before endoscopic resection or any other treatment and were centrifuged at 3,000 × g for 15 min at 4°C for the isolation of plasma. All participants had signed the informed consent, and this study was approved by the ethics committee of Beijing Friendship Hospital.

Highly Abundant Protein Depletion

All plasma samples for SWATH-MS analysis were processed using Pierce Top 12 Abundant Protein Depletion Spin Columns (#85164, Thermo Fisher Scientific, Waltham, MA, USA) according to the

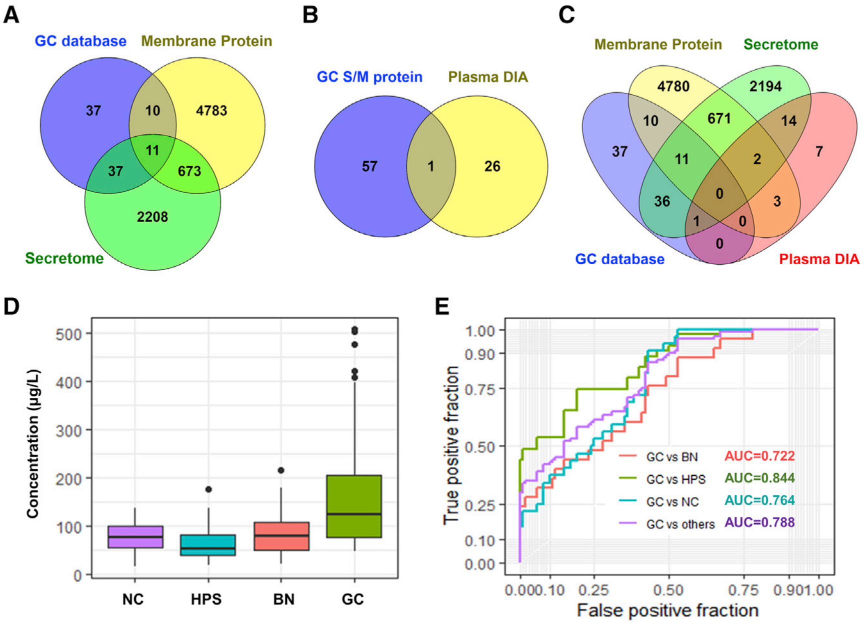

Figure 6. Identification and ELISA Verification of CHI3L1 as a Highly Elevated Secreted Protein in GC Plasma

(A) A Venn diagram showed that 58 out of 94 database-derived upregulated genes were secretome or membrane associated. (B) There was only 1 gene, CHI3L1, characterized as upregulated in both the database-derived list and SWATH-MS. (C) An overall Venn diagram showed the overlap among GC dataset consensus DEGs, SWATH-MS DIA-identified DEGs in GC plasma, membrane protein lists, and secretome. (D) A boxplot showed different CHI3L1 levels in plasma of patients with GC, patients with BN, patients with HPs, and NCs by ELISA. (E) ROC plots exhibited the efficiencies of CHI3L1 in distinguishing GC patients from those with precancerous BN, those with HPs, NCs, and all of the above.

manufacturer's protocol. Briefly, 10 μL plasma was added to the column and well mixed for 60 min; then, the eluent collected by centrifugation at 1,000 × *g* for 2 min was frozen at −80°C until further use.

Protein Digestion and Pre-separation

The depleted sample was digested using the filter-aided sample preparation (FASP) method.[42] Briefly, disulfide bonds were broken and blocked using 10 mM dithiothreitol (DTT) and 50 mM iodoacetamide (IAA), and then proteins were transferred to a 10-kDa filter and cleaned sequentially using 8 M urea and 50 mM NH_4HCO_3 at 13,000 × *g*, at 20°C. Trypsin was added to each sample at 1:50 (mass:mass) in 50 mM NH_4HCO_3 at 37°C for 16 h. A mixture sample was made of an equal amount of digested peptides from each sample and separated into 3 fractions using a modified high-pH RP method.[43] Briefly, a homemade C18 stage tip was cleaned using 80% acetonitrile (ACN)/H_2O after activation with methanol. Then, the stage tip was equilibrated with ammonium hydroxide (pH 10) before peptides were loaded onto the stage tip. A series of ACN/ammonium hydroxide (pH 10) buffer concentrations—6%, 9%, 12%, 15%, 18%, 21%, 25%, 30%, 35%, and 50%—was used to elute peptides into 10 fractions and then combined into 3 fractions.

DDA sample preparation and MS procedures

For generation of the spectral library, the pre-separated 3 fractions were acquired with the DDA mode using Orbitrap Fusion Lumos (#IQLAAEGAAPFADBMBHQ; Thermo Fisher Scientific, San Jose, CA, USA). The peptide mixture was separated on the Easy-nLC 1200 System (#LC140, Thermo Fisher Scientific, San Jose, CA, USA) using a homemade C18 column (3 μm, 75 μm × 15 cm) at a flow rate of 600 nL/min. A 120-min linear gradient was set as follows: 5% reagent B (0.1% FA in 80% ACN/H_2O)/95% reagent A (0.1% FA in H_2O) to 10% reagent B in 13 min, 10% reagent B to 30% reagent B in 80 min, 30% reagent B to 45% reagent B in 20 min, 45% reagent B to 95% reagent B in 1 min, and stayed 6 min for 95% B. For the data acquisition, a top 20 scan mode with MS1 scan range m/z 400–1,200, was used. Other parameters were set as follows: MS1 and MS2 resolution was set to 120K and 30K, respectively; AGC for MS1 and MS2 was 4e5 and 1e5; isolation window was 2.0 Th; NCE was 32; and dynamic exclusion time was 15 s.

DIA sample preparation and MS procedures

Each sample with an addition of the same amount of iRT (Biognosys)[44] was analyzed with the DIA method. For DIA acquisition, the method consisted of one full MS1 scan with a resolution set at 120K using AGC of 4e5 and a maximum injection time of 50 ms. A sequential 29 isolation mass window was set as follows: for m/z 400–800, the mass isolation window was set to 20 Th; for m/z 800–1,000, the mass isolation window was set to 40 Th; and for m/z 1,000–1,200, the mass isolation window was set to 50 Th. Each DIA MS2 spectrum was acquired using a resolution of 30K, AGC was set to 1e5, maximum injection time was 50 ms, and collision energy was set to NCE 35. All the LC conditions were exactly the same as for DDA listed earlier.

Spectral Library Generation

DDA raw files were searched against a UniProt protein database of *Homo sapiens* using Proteome Discoverer v.2.1. The protein sequence was appended with the iRT fusion protein sequence. A search engine of SequestHT was used with the following searching parameter: enzyme of trypsin with maximum number of 2 missed cleavages; precursor and fragment ion mass tolerance was set to 10 ppm and 0.02 Da; variable modification was set to oxidation of M and deamidation of N, Q; and fixed modification was set to carbamidomethylation of C. An algorithm of Percolator[45] was used to keep peptide FDR less than 1% and the q value threshold used for protein identification was less than 0.01. Search results of DDA using Proteome Discoverer 2.1 were transferred into a spectral library using Spectronaut 10 (Biognosys). Only high confidence of peptide was used for the generation of the spectral library.

Table 1. Clinical Characteristics of Patients with GC, BN, and HPS and of NC

Clinical Characteristics		Non-Cancerous Controls (n = 100)		
	GC (n = 100)	BN (n = 25)	HPS (n = 43)	NC (n = 32)
Age (mean ± SD)	54.99 ± 10.24	55.88 ± 12.90	57.30 ± 11.14	53.12 ± 15.01
Gender				
Male	76	11	12	24
Female	24	14	31	8
Clinical Stage				
I	79	–	–	–
II/III	21	–	–	–
Tumor Location				
Cardia/Fundus	20	5	18	–
Body	20	7	17	–
Antrum	49	13	7	–
Muti-site	11	0	1	–
Paris Subtype				
Ip	0	2	5	–
Is	0	12	10	–
Isp	3	11	28	–
II	97	0	0	–
Lesion size (mean ± SD)	3.60 ± 1.79	1.18 ± 0.79	0.95 ± 0.63	–
CHI3L1 level (μg/L) (mean ± SD)	162.59 ± 112.78	87.40 ± 50.33	64.50 ± 35.45	76.21 ± 29.49

Fragment ions within the mass range of m/z 300–1,800 were kept, and peptides less than 3 fragment ions were removed.

DIA Data Analysis

Search results of data-dependent acquisition using Proteome Discoverer v.2.1 were transferred into a spectral library using Spectronaut 10 (Biognosys, Schlieren, Switzerland). The following settings were applied in Spectronaut 10.0: peak detection, dynamic iRT, enabled; correction factor 1; dynamic score refinement and MS1 scoring, enabled; interference correction and cross run normalization (total peak area), enabled. The number of fragment ions was defined in the spectral library, and all were required for identification and quantification. Spectronaut utilizes the spiked-in iRT peptides for m/z and retention time calibration. For our dataset, the m/z tolerance was in the range of 4 ppm, and the median retention time extraction window was 8 min. All results were filtered by a Q value of 0.01 (equals an FDR of 1% on peptide level). All other settings were set to default. Protein intensity was calculated by summing the peptide peak areas (sum of fragment ion peak areas as calculated by Spectronaut) of each protein from the Spectronaut output file.

Bioinformatic Analysis and Visualization

TCGA and four GEO datasets (GEO: GSE54129, GSE29272, GSE26942, and GSE37023) were downloaded for analysis. DEGs were identified by the GEO2R platform, with all parameters set at default.[46] The annotation of the given gene list was performed with ClueGO function analysis.[47] ClueGO function analysis was performed based on kappa score grouping; at global, medium, and detailed levels, separately, with all other parameters set at default. Redundant groups with >50.0% overlap were merged as one GO term. Packages including pheatmap, VennDiagram, and plotROC of R v.3.3.1 were used for results visualization.

ELISA

CHI3L1 levels in plasma specimens were measured using a commercially available ELISA kit (#CSB-E13608h, CUSABIO, Wuhan, China) according to the manufacturer's protocol. Plasma samples were diluted 1:5 before analysis. Each sample was detected in duplicate, and the mean value was used as the final readout. A sample with a coefficient of variation (CV)% > 8% was re-tested to ensure data reliability.

Statistical Analysis

We used Student's t test, χ2 test, or one-way ANOVA to evaluate the association between CHI3L1 level and clinicopathological characteristics. p values < 0.05 were considered statistically significant. ROC curves were drawn to evaluate the predictive accuracy of the CHI3L1 level for GC, which was quantified by the AUC. R v.3.3.1 was used for data analysis, and ggplot2 was used for visualization of results.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omto.2020.03.020.

AUTHOR CONTRIBUTIONS

L.M. and S. Zhang designed the study. L.M., S. Zhu, and R.W. performed all the experiments. L.M. and P.L. performed the statistical analyses and interpreted the data. Y.Z., S.L., and P.L. helped to collect clinical samples and data. L.M. drafted the manuscript. All authors contributed to the final version of the manuscript and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank Ms. Yun Zhang, Ms. Yilin Chen, and Ms. Ping Zhang from the Clinical Data and Biobank Resource of Beijing Friendship Hospital for their dedicated help in clinical sample collection. We also thank the high-level engineers from Beijing Biolino Technologies for their professional technical support of SWATH-MS.

This study was funded by the Beijing Science and Technology Nova Program (Z191100001119128); the Beijing Municipal Administration of Hospitals' Youth Programme (QML20180108); the National Natural Science Foundation of China (81702314); the Beijing Municipal Science and Technology Project (Z191100006619081); the Funding Program for Excellent Talents of Beijing (2017000021469G212);

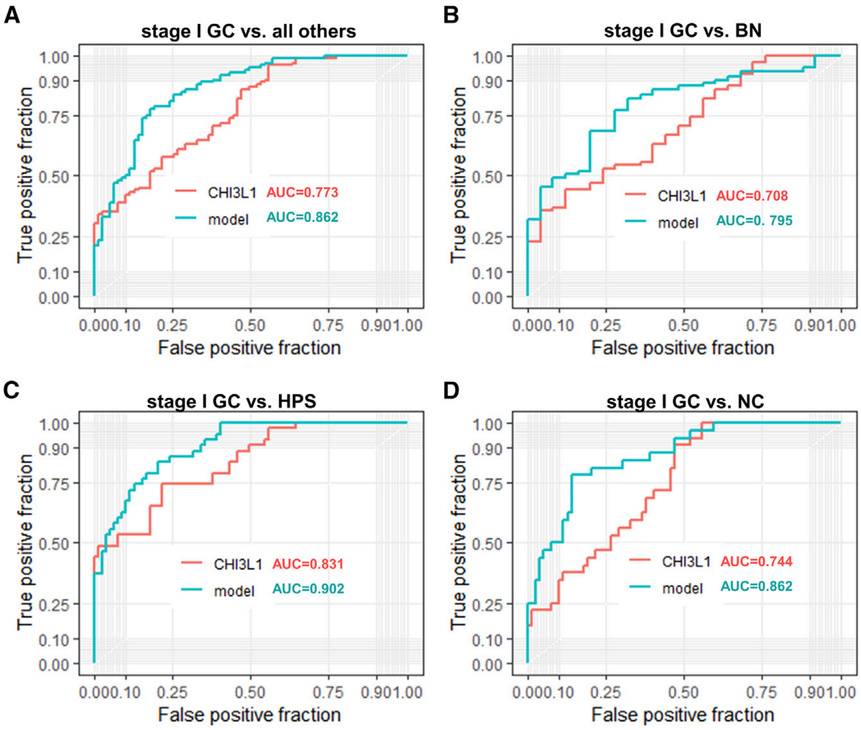

Figure 7. Integrating CHI3L1 with Clinical Factors Could Well Distinguish Patients with Early-Stage Endoscopically Resectable GC from Others

(A) The ROC plot of stage-I GC patients versus all other participants. (B) The ROC plot of stage-I GC patients versus BN patients. (C) The ROC plot of stage-I GC patients versus those with HPs. (D) The ROC plot of stage-I GC patients versus NCs.

and The Digestive Medical Coordinated Development Center of Beijing Municipal Administration of Hospitals (XXZ0201). The study sponsors had no role in the design and preparation of this manuscript.

REFERENCES

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.
2. Siegel, R.L., Miller, K.D., and Jemal, A. (2017). Cancer Statistics, 2017. CA Cancer J. Clin. *67*, 7–30, https://doi.org/10.3322/caac.21387.
3. Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. (2016). Cancer statistics in China, 2015. CA Cancer J. Clin. *66*, 115–132.
4. Li, F., Zhang, R., Liang, H., Liu, H., Quan, J., and Zhao, J. (2012). The pattern of lymph node metastasis and the suitability of 7th UICC N stage in predicting prognosis of remnant gastric cancer. J. Cancer Res. Clin. Oncol. *138*, 111–117.
5. Salem, A., Hashem, S., Mula-Hussain, L.Y., Mohammed, I., Nour, A., Shelpai, W., Daoud, F., Morcos, B., Yamin, Y., Jaradat, I., et al. (2012). Management strategies for locoregional recurrence in early-stage gastric cancer: retrospective analysis and comprehensive literature review. J. Gastrointest. Cancer *43*, 77–82.
6. Zong, L., Abe, M., Seto, Y., and Ji, J. (2016). The challenge of screening for early gastric cancer in China. Lancet *388*, 2606.
7. Song, W.C., Qiao, X.L., and Gao, X.Z. (2015). A comparison of endoscopic submucosal dissection (ESD) and radical surgery for early gastric cancer: a retrospective study. World J. Surg. Oncol. *13*, 309.
8. Kato, M. (2005). Endoscopic submucosal dissection (ESD) is being accepted as a new procedure of endoscopic treatment of early gastric cancer. Intern. Med. *44*, 85–86.
9. Allemani, C., Weir, H.K., Carreira, H., Harewood, R., Spika, D., Wang, X.S., Bannon, F., Ahn, J.V., Johnson, C.J., Bonaventure, A., et al.; CONCORD Working Group (2015). Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet *385*, 977–1010.
10. Oh, S., Kim, N., Yoon, H., Choi, Y.J., Lee, J.Y., Park, K.J., Kim, H.J., Kang, K.K., Oh, D.H., Seo, A.Y., et al. (2013). Risk factors of atrophic gastritis and intestinal metaplasia in first-degree relatives of gastric cancer patients compared with age-sex matched controls. J. Cancer Prev. *18*, 149–160.
11. Sipponen, P., Kekki, M., and Siurala, M. (1988). Increased risk of gastric cancer in males affects the intestinal type of cancer and is independent of age, location of the tumour and atrophic gastritis. Br. J. Cancer *57*, 332–336.
12. Shin, C.M., Kim, N., Yang, H.J., Cho, S.I., Lee, H.S., Kim, J.S., Jung, H.C., and Song, I.S. (2010). Stomach cancer risk in gastric cancer relatives: interaction between Helicobacter pylori infection and family history of gastric cancer for the risk of stomach cancer. J. Clin. Gastroenterol. *44*, e34–e39.
13. Zullo, A., Hassan, C., and Morini, S. (2002). Helicobacter pylori infection and the development of gastric cancer. N. Engl. J. Med. *346*, 65–67.
14. Yoo, M.W., Park, J., Han, H.S., Yun, Y.M., Kang, J.W., Choi, D.Y., Lee, J.W., Jung, J.H., Lee, K.Y., and Kim, K.P. (2017). Discovery of gastric cancer specific biomarkers by the application of serum proteomics. Proteomics *17*, 1600332.
15. Alimirzaie, S., Bagherzadeh, M., and Akbari, M.R. (2019). Liquid biopsy in breast cancer: A comprehensive review. Clin. Genet. *95*, 643–660.
16. Waldron, D. (2016). Cancer genomics: A nucleosome footprint reveals the source of cfDNA. Nat. Rev. Genet. *17*, 125.
17. Wu, W., Juan, W.C., Liang, C.R., Yeoh, K.G., So, J., and Chung, M.C. (2012). S100A9, GIF and AAT as potential combinatorial biomarkers in gastric cancer diagnosis and prognosis. Proteomics Clin. Appl. *6*, 152–162.
18. Wu, W., Yong, W.W., and Chung, M.C. (2016). A simple biomarker scoring matrix for early gastric cancer detection. Proteomics *16*, 2921–2930.
19. Li, Z., Zuo, X.L., Li, C.Q., Zhou, C.J., Liu, J., Goetz, M., Kiesslich, R., Wu, K.C., Fan, D.M., and Li, Y.Q. (2013). In vivo molecular imaging of gastric cancer by targeting MG7 antigen with confocal laser endomicroscopy. Endoscopy *45*, 79–85.

20. Xu, B., Li, X., Yin, J., Liang, C., Liu, L., Qiu, Z., Yao, L., Nie, Y., Wang, J., and Wu, K. (2015). Evaluation of 68Ga-labeled MG7 antibody: a targeted probe for PET/CT imaging of gastric cancer. Sci. Rep. *5*, 8626.

21. Anjo, S.I., Santa, C., and Manadas, B. (2017). SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. Proteomics *17*, 1600278.

22. Luo, Y., Mok, T.S., Lin, X., Zhang, W., Cui, Y., Guo, J., Chen, X., Zhang, T., and Wang, T. (2017). SWATH-based proteomics identified carbonic anhydrase 2 as a potential diagnosis biomarker for nasopharyngeal carcinoma. Sci. Rep. *7*, 41191.

23. Hou, G., Lou, X., Sun, Y., Xu, S., Zi, J., Wang, Q., Zhou, B., Han, B., Wu, L., Zhao, X., et al. (2015). Biomarker Discovery and Verification of Esophageal Squamous Cell Carcinoma Using Integration of SWATH/MRM. J. Proteome Res *14*, 3793–3803.

24. Liu, Y., Chen, J., Sethi, A., Li, Q.K., Chen, L., Collins, B., Gillet, L.C., Wollscheid, B., Zhang, H., and Aebersold, R. (2014). Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. Mol. Cell. Proteomics *13*, 1753–1768.

25. Gao, Y., Wang, X., Sang, Z., Li, Z., Liu, F., Mao, J., Yan, D., Zhao, Y., Wang, H., Li, P., et al. (2017). Quantitative proteomics by SWATH-MS reveals sophisticated metabolic reprogramming in hepatocellular carcinoma tissues. Sci. Rep. *7*, 45913.

26. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209.

27. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

28. Coffman, F.D. (2008). Chitinase 3-Like-1 (CHI3L1): a putative disease marker at the interface of proteomics and glycomics. Crit. Rev. Clin. Lab. Sci. *45*, 531–562.

29. Ngernyuang, N., Shao, R., Suwannarurk, K., and Limpaiboon, T. (2018). Chitinase 3 like 1 (CHI3L1) promotes vasculogenic mimicry formation in cervical cancer. Pathology *50*, 293–297.

30. Fan, J.T., Si, X.H., Liao, Y., and Shen, P. (2013). The diagnostic and prognostic value of serum YKL-40 in endometrial cancer. Arch. Gynecol. Obstet. *287*, 111–115.

31. Qiu, Q.C., Wang, L., Jin, S.S., Liu, G.F., Liu, J., Ma, L., Mao, R.F., Ma, Y.Y., Zhao, N., Chen, M., and Lin, B.Y. (2018). CHI3L1 promotes tumor progression by activating TGF-β signaling pathway in hepatocellular carcinoma. Sci. Rep. *8*, 15029.

32. Wang, J., Gao, F., Mo, F., Hong, X., Wang, H., Zheng, S., and Lin, B. (2009). Identification of CHI3L1 and MASP2 as a biomarker pair for liver cancer through integrative secretome and transcriptome analysis. Proteomics Clin. Appl. *3*, 541–551.

33. Libreros, S., Garcia-Areas, R., Shibata, Y., Carrio, R., Torroella-Kouri, M., and Iragavarapu-Charyulu, V. (2012). Induction of proinflammatory mediators by CHI3L1 is reduced by chitin treatment: decreased tumor metastasis in a breast cancer model. Int. J. Cancer *131*, 377–386.

34. Liu, X., Zhang, Y., Zhu, Z., Ha, M., and Wang, Y. (2014). Elevated pretreatment serum concentration of YKL-40: an independent prognostic biomarker for poor survival in patients with colorectal cancer. Med. Oncol. *31*, 85.

35. Bi, J., Lau, S.H., Lv, Z.L., Xie, D., Li, W., Lai, Y.R., Zhong, J.M., Wu, H.Q., Su, Q., He, Y.L., et al. (2009). Overexpression of YKL-40 is an independent prognostic marker in gastric cancer. Hum. Pathol. *40*, 1790–1797.

36. (2015). Interactive human protein atlas launches. Cancer Discov. *5*, 339.

37. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

38. Lin, J., Chen, J., and Liu, C. (2015). AGT M235T variant is not associated with risk of cancer. J. Renin Angiotensin Aldosterone Syst. *16*, 448–452.

39. Cine, N., Baykal, A.T., Sunnetci, D., Canturk, Z., Serhatli, M., and Savli, H. (2014). Identification of ApoA1, HPX and POTEE genes by omic analysis in breast cancer. Oncol. Rep. *32*, 1078–1086.

40. Sun, H.M., Mi, Y.S., Yu, F.D., Han, Y., Liu, X.S., Lu, S., Zhang, Y., Zhao, S.L., Ye, L., Liu, T.T., et al. (2016). SERPINA4 is a novel independent prognostic indicator and a potential therapeutic target for colorectal cancer. Am. J. Cancer Res. *6*, 1636–1649.

41. Kamba, A., Lee, I.A., and Mizoguchi, E. (2013). Potential association between TLR4 and chitinase 3-like 1 (CHI3L1/YKL-40) signaling on colonic epithelial cells in inflammatory bowel disease and colitis-associated cancer. Curr. Mol. Med. *13*, 1110–1121.

42. Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

43. Dimayacyac-Esleta, B.R., Tsai, C.F., Kitata, R.B., Lin, P.Y., Choong, W.K., Lin, T.D., Wang, Y.T., Weng, S.H., Yang, P.C., Arco, S.D., et al. (2015). Rapid High-pH Reverse Phase StageTip for Sensitive Small-Scale Membrane Proteomic Profiling. Anal. Chem. *87*, 12016–12023.

44. Bruderer, R., Bernhardt, O.M., Gandhi, T., and Reiter, L. (2016). High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. Proteomics *16*, 2246–2256.

45. Spivak, M., Weston, J., Bottou, L., Käll, L., and Noble, W.S. (2009). Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J. Proteome Res. *8*, 3737–3745.

46. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

47. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.