EBioMedicine 74 (2021) 103714

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Original Research

Proteomic profiling identifies signatures associated with progression of precancerous gastric lesions and risk of early gastric cancer

Xue Li[a,†], Nai-Ren Zheng[b,†], Lin-Heng Wang[c,†], Zhong-Wu Li[d], Zong-Chao Liu[a], Hua Fan[a], Yi Wang[b], Jin Dai[a], Xiao-Tian Ni[b], Xin Wei[b], Ming-Wei Liu[b], Kai Li[b], Zhe-Xuan Li[a], Tong Zhou[a], Yang Zhang[a], Jing-Ying Zhang[a], Gaohaer Kadeerhan[a], Sha Huang[a], Wen-Hui Wu[a], Wei-Dong Liu[e], Xiu-Zhen Wu[f], Lan-Fu Zhang[f], Jian-Ming Xu[g], Markus Gerhard[h], Wei-Cheng You[a,h], Kai-Feng Pan[a,h,3], Wen-Qing Li[a,h,2,*], Jun Qin[b,3]

[a] *Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Cancer Epidemiology, Peking University Cancer Hospital & Institute, Beijing 100142, China*
[b] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China*
[c] *Department of Gastroenterology, Second Clinical Medical College of Beijing University of Chinese Medicine (Dongfang Hospital), Beijing 100078, China*
[d] *Department of Pathology, Peking University Cancer Hospital & Institute, Beijing 100142, China*
[e] *Linqu County Public Health Bureau, Shandong 262600, China*
[f] *Linqu County People's Hospital, Shandong 262600, China*
[g] *Department of Gastrointestinal Oncology, The Fifth Medical Center, General Hospital of PLA, Beijing 100071, China*
[h] *PYLOTUM Key joint laboratory for upper GI cancer, Technische Universität München/Peking University Cancer Hospital & Institute, Munich/Beijing, Germany/China*

ARTICLE INFO

Article History:
Received 23 June 2021
Revised 9 November 2021
Accepted 9 November 2021
Available online xxx

Keywords:
Proteomics
Precancerous gastric lesions
Gastric cancer
Biomarker

ABSTRACT

Background: Molecular features underlining the multistage progression of gastric lesions and development of early gastric cancer (GC) are poorly understood, restricting the ability to GC prevention and management.
Methods: We portrayed proteomic landscape and explored proteomic signatures associated with progression of gastric lesions and risk of early GC. Tissue proteomic profiling was conducted for a total of 324 subjects. A case-control study was performed in the discovery stage (n=169) based on populations from Linqu, a known high-risk area for GC in China. We then conducted two-stage validation, including a cohort study from Linqu (n = 56), with prospective follow-up for progression of gastric lesions (280–473 days), and an independent case-control study from Beijing (n = 99).
Findings: There was a clear distinction in proteomic features for precancerous gastric lesions and GC. We derived four molecular subtypes of gastric lesions and identified subtype-S4 with the highest progression risk. We found 104 positively-associated and 113 inversely-associated proteins for early GC, with APOA1BP, PGC, HPX and DDT associated with the risk of gastric lesion progression. Integrating these proteomic signatures, the ability to predict progression of gastric lesions was significantly strengthened (areas-under-the-curve=0.88 (95%CI: 0.78–0.99) vs. 0.56 (0.36–0.76), Delong's P = 0.002). Immunohistochemistry assays and examination at mRNA level validated the findings for four proteins.
Interpretation: We defined proteomic signatures for progression of gastric lesions and risk of early GC, which may have translational significance for identifying particularly high-risk population and detecting GC at an early stage, improving potential for targeted GC prevention.
Funding: The funders are listed in the Acknowledgement.

1. Introduction

Gastric cancer (GC) is a leading cause of cancer-related deaths threatening public health worldwide [1]. Most patients are diagnosed at an advanced stage with limited treatment options and poor prognosis whereas efficient prevention is still a bottleneck. The current

* Corresponding author.
E-mail addresses: pankaifeng2002@yahoo.com (K.-F. Pan), wenqing_li@bjmu.edu.cn (W.-Q. Li), jqin1965@126.com (J. Qin).
† X. L., N.-R. Z. and L.-H. W. contributed equally to this paper.
[2] Primary corresponding author
[3] Co-corresponding author

https://doi.org/10.1016/j.ebiom.2021.103714

Research in context

Evidence before this study

The development of gastric cancer is preceded by a prolonged multi-stage precancerous process, but molecular signatures underling the dynamic changes of gastric lesions and development of early gastric cancer are poorly understood, restricting the ability to its prevention and management. Identifying molecular signatures and biomarkers associated with the progression of precancerous gastric lesions and development of early gastric cancer is of great public health significance for lowering gastric cancer burden worldwide.

Added value of this study

We portrayed a proteomic landscape for gastric lesions of different stages and gastric cancer and found four proteomic subtypes of gastric lesions with different progression potential based on prospective follow-up of subjects, revealing molecular heterogeneity of gastric lesions beyond cellular morphology. Four proteins, including APOA1BP, PGC, HPX and DDT, were validated to be associated with the risk of gastric lesion progression and early gastric cancer. Integrating these proteomic signatures, a risk prediction model significantly improved the ability to predict progression of gastric lesions.

Implications of all the available evidence

This is the first study that comprehensively explored the dynamic changes of proteome and protein signatures in the evolution of gastric lesions and early gastric cancer. The defined proteomic signatures for gastric lesion progression and risk of early gastric cancer may have translational significance for identifying populations at a particular high-risk for gastric cancer and detecting gastric cancer at a very early stage, improving the ability to its targeted prevention and management.

GC prevention program relies mostly on gastroscopy screening but concerns over its invasive approach, requirements on skilled endoscopists and pathologists, and a high cost have been tangled for developing countries [2].

Our endoscopy-based cohort study in Linqu county, Shandong province of China, an area with one of the highest mortality of GC worldwide [3–5], revealed that the risk of invasive GC was remarkably increased by gastric histopathology, with an odds ratio (OR) of 17.4 for mild intestinal metaplasia (IM), 29.3 for moderate-to-severe IM or low-grade intraepithelial neoplasia (LGIN), and 104.2 for high-grade intraepithelial neoplasia (HGIN), compared with superficial gastritis (SG) or chronic atrophic gastritis (CAG) [4], representing a multistage gastric carcinogenesis process. Even so, the progression of advanced gastric lesions only occurs in a small proportion of individuals [5]. Identifying populations vulnerable for progression of gastric lesions and detecting GC at an early stage is warranted so that appropriate prevention strategies can be implemented. However, other than *Helicobacter pylori* (*H. pylori*) [5], risk factors of GC are largely unknown and the etiology behind the progression of gastric lesions to GC remains to be clarified. Established molecular signatures underlying gastric lesion progression and GC development are very limited and well-performed biomarkers are highly needed particularly based on prospective studies.

Although previous studies have identified clinically relevant molecular subtypes and genomic alterations of GC [6,7], related evidence for precancerous gastric lesions and early-stage GC has been sparse. Realizing proteins as the “executioners of life” that carry out the functions of genes and determine phenotype, some proteomic and proteogenomic studies have been conducted on GC subtypes for new therapeutic targets [8,9] and predicting cancer prognosis [10], focusing mostly on diffuse type GC. A few proteomic studies have involved gastric lesions [11–18], such as gastritis and gastric ulcers, but only used these lesions as controls to investigate the proteomic changes in GC. Several studies investigated proteomic changes in IM [19,20] or dysplasia [21,22] but were limited by either lack of controlling for possible confounders and multiple comparison corrections or small sample size. No studies have prospectively examined the proteomic changes along the cascade of evolution of precancerous gastric lesions and GC development.

By in-depth proteomic profiling of 324 subjects, we portrayed a proteomic landscape of precancerous gastric lesions and GC and explored proteomic signatures, including proteomic subtypes and individual proteins associated with the progression of precancerous gastric lesions and risk of early GC. Our ongoing work in Linqu county, an established high-risk area for GC, has provided unique opportunity allowing for addressing our goals with a longitudinal study design. We further validated the findings based on an independent cohort with a ‘normal’ risk of GC from Beijing and developed immunohistochemistry (IHC) assays for key proteins to facilitate future preventive and clinical translation.

2. Methods

2.1. Study subjects

Our studies had a discovery stage and two-stage validation involving a total of 324 subjects. The discovery stage was designed as an unmatched case-control study and involved 169 subjects, including 111 with gastric lesions at various stages and 58 with invasive GC (Fig. 1a and Table S1). The case-control study defined the well-recognized mild gastric lesions (SG or CAG) as control group, and had advanced gastric lesions (IM or LGIN) and GC as two case groups. Individuals of gastric lesions were enrolled based on the National Upper Gastrointestinal Cancer Early Detection (UGCED) Program in Linqu, Shandong province of China, where most GC cases are of intestinal type. From Nov 22, 2018 to Dec 07, 2018, 231 Linqu residents aged 40-69 years were randomly selected to undertake gastroscopies in the UGCED program. Controls would therefore be a representative sample of individuals in Linqu with mild gastric lesions. In this program, cases diagnosed with HGIN would be classified as ‘very early’ GC and treated immediately along with invasive GC cases. Gastric biopsies were obtained for pathological diagnosis and an extra biopsy was taken from lesser curvature of antrum for proteomic assays. We did not find any individual with completely normal gastric mucosa. To avoid systemic effects on proteomic profiles caused by severer gastric lesions at other gastric mucosa sites, we only included subjects that had the most severe histology of gastric mucosa at lesser curvature of antrum. A total of 111 subjects remained, including 33 SGs, 19 CAGs, 56 IMs, 3 LGINs, and none with HGIN or invasive GC. As we did not identify GCs among 231 Linqu residents, we procured 58 invasive GC cases (19 of clinical stage II-III, 38 of stage IV, and 1 of unknown stage, UICC&AJCC TNM classification, 8th edition; 17 of intestinal type, 16 of diffuse type, 20 of mixed type and 5 unknown) from the Fifth Medical Center, General Hospital of PLA for analyses. All cancer tissues were collected before chemotherapy. Details of these GC cases have been described previously [10].

We conducted two-stage validation enrolling 155 subjects, including 108 with gastric lesions and 47 diagnosed with early GC based on two independent cohorts from Linqu and Beijing (Fig.1a and Table S1). All GCs were of intestinal type. For Linqu validation set, a cohort study was conducted with prospective follow-up. This set involved 56 subjects, including 39 subjects diagnosed with LGIN or less severe gastric lesions from Sep 27, 2017 to May 22, 2018 and

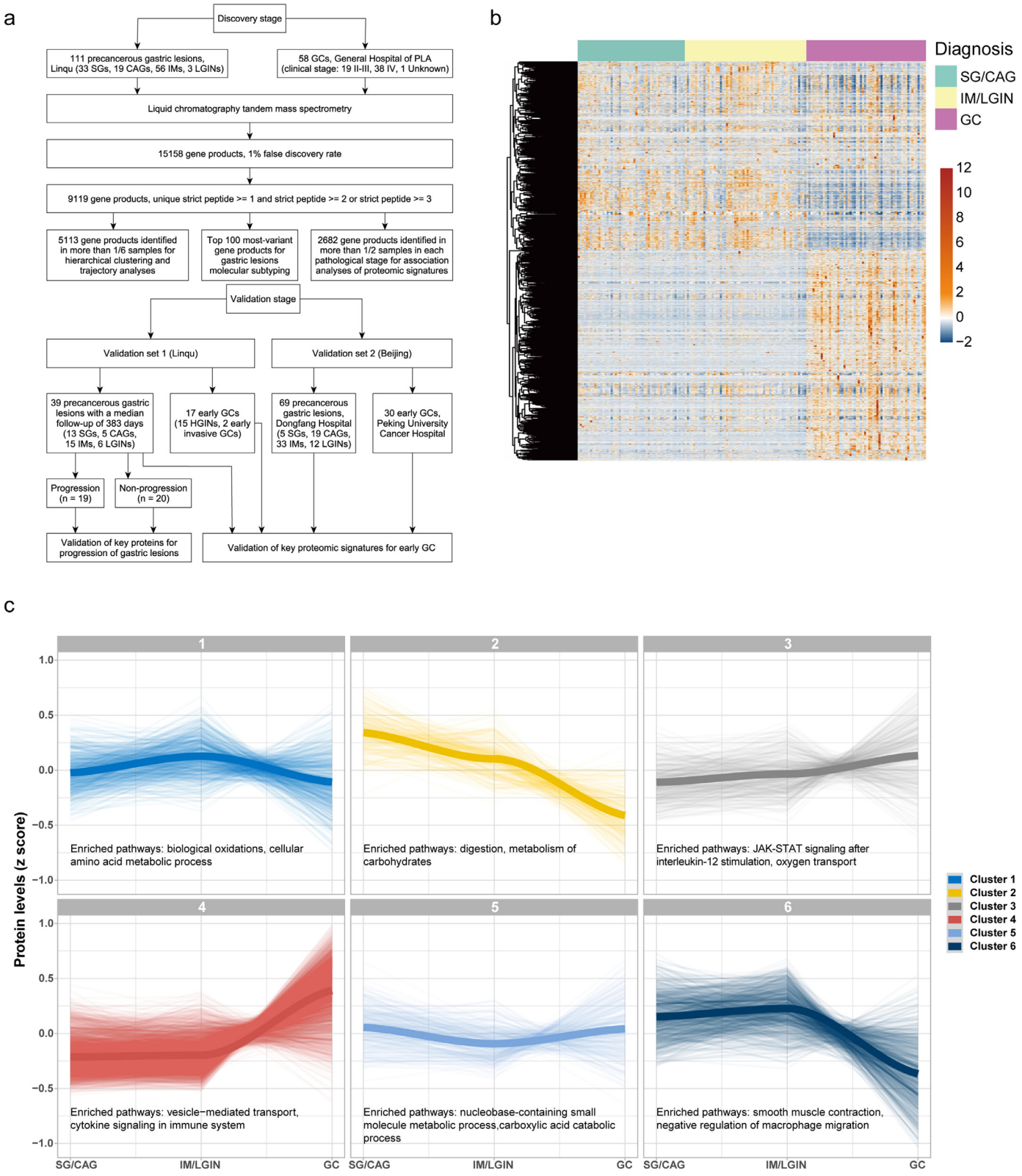

Fig. 1. Outline of work flow and proteomic landscape of precancerous gastric lesions and GC. **a.** Study design and flow chart. **b.** Hierarchical clustering of protein profiles in three pre-defined subject categories: mild gastric lesions (SG/CAG), advanced gastric lesions (IM/LGIN) and GC, with each column denoting one subject and each row denoting one protein. **c.** Six aggregated protein clusters with similar trajectories from mild (SG/CAG) to advanced (IM/LGIN) gastric lesions and then to GC. CAG, chronic atrophic gastritis; GC, gastric cancer; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; SG, superficial gastritis.

17 subjects diagnosed with early GC, including HGIN (n = 15) and early-stage invasive GC (n = 2) from Sep 27, 2017 to Apr 8, 2019, based on the National UGCED Program in Linqu. These 39 subjects of gastric lesions were prospectively followed for 280 to 473 days (median 383 days), with endoscopic examinations conducted again at endpoint. We assessed the evolution of gastric lesions at the same

biopsy site during follow-up (Table S2). Each subject was assigned a histology severity score at baseline and endpoint respectively. A subject was classified to have progression of gastric lesions if the endpoint score was higher than the baseline. During endpoint endoscopy, 21 out of 39 subjects also had a biopsy specifically taken for proteomic profiling at the same mucosa site with the baseline.

To test the extrapolation of findings to populations with a relatively low risk for GC, we then had another unmatched case-control study (the same approach defining control and case groups with the discovery stage) using an independent validation set, with 69 subjects of gastric lesions (5 SGs, 19 CAGs, 33 IMs, 12 LGINs) from Dongfang Hospital and 30 of early GC from Peking University Cancer Hospital in Beijing. The controls were randomly selected from individuals diagnosed with SG or CAG between Mar 7 and Dec 5 in 2019.

For the validation stages, tissues were collected from the biopsy sites which had the most severe histology of gastric mucosa for proteomic profiling, all from lesser curvature of antrum or angulus to avoid systemic bias of different mucosal sites on proteomic profiles.

All biospecimens were prepared and deposited following the same processing and preservation methods. A 5 ml blood sample was collected from each subject and *H. pylori* antibody assays were used for determining *H. pylori* infection status for subjects from Linqu [23].

2.2. Ethics

The study was approved by the Institutional Review Boards of Peking University Cancer Hospital (approval No. 2018KT117) and Dongfang Hospital (JDF-IRB-B20190308). All participants provided written informed consent.

2.3. Gastroscopy and histopathology

Gastroscopic examinations were conducted by two experienced gastroenterologists using video endoscopes (Olympus). Biopsies were reviewed blindly by two pathologists according to the criteria proposed by the Chinese Association of Gastric Cancer and Updated Sydney System [24,25]. Each biopsy was given a diagnosis based on the most severe histology. Each subject was assigned a 'global' diagnosis based upon the most severe diagnosis among all biopsies.

2.4. Liquid chromatography-tandem mass spectrometry assay (LC-MS/MS)

For proteomic profiling, LC-MS/MS assay was performed on a high-resolution quadrupole Orbitrap Fusion and Orbitrap Fusion Lumos MS coupled with an Easy-nLC 1000 nanoflow LC system (Thermo Fisher Scientific) in a data-dependent mode. Tissue samples were minced and lysed in a buffer (8 M Urea, 100 mM Tris-HCl at pH 8.5) and denatured at 95 ℃ for 5 min, followed by sonication in an ice-water bath for 5 min. Extracts from each sample (50 μg protein) were digested using trypsin and evaporated to dryness. After being redissolved in 0.2% formic acid, dried protein digests were loaded to LC-MS/MS with a gradient of 5−35% mobile phase A (H_2O: FA = 99.8:0.2) and B (CAN:FA = 99.8:0.2) for 141 min then up to 95% in 1min and eluted for 9min. Flow rate was kept at 600 nL/min and the column temperature was maintained at 60 °C. Tryptic digestions of 293T cell lysate were used as quality control (QC) samples. Pairwise Spearman's correlation coefficients (r) were calculated for all QC runs through corrplot package v.0.84 in R.

2.5. Data processing and protein quantification

We followed well-established approaches for data processing and protein quantification. The Firmiana pipeline was used to process raw mass spectra [26]. Mascot search engine (Matrix Science, version 2.3) was used to identify proteins against the NCBI human Refseq protein database (version 04/07/2013) with mass tolerances of 20ppm for precursors and 0.5Da for product ions. Up to 2 missed cleavages were allowed and 1% false discovery rate (FDR) on the peptide level was considered acceptable (n = 15158 gene products). Peptides with Mascot ions score higher than 20 were defined as strict peptides. Strict peptides belonging to only one protein were further classified as unique strict peptides. Proteins with at least 1 unique strict peptide and 2 strict peptides, and proteins with at least 3 strict peptides are considered of high reliability (n = 9119). Among them, 5113 proteins were identified in more than 1/6 samples and 2682 were identified in more than 1/2 samples.

A label-free intensity-based absolute quantification (iBAQ) approach was used to quantify protein abundance [27]. The iBAQ values were then converted to intensity-based fraction of total (iFOT), calculated as the iBAQ of each protein divided by the sum of iBAQs of all proteins in the sample and multiplied by 10^5 to ease the visualization of low abundant proteins [28]. The normalized and $\log_{10}$ transformed iFOT values were plotted for each sample to show consistency of data quality.

2.6. Bioinformatics and statistical analysis

All analyses were conducted using R (v.3.6.0) unless otherwise noted. In Linqu, SG represents the least abnormal type of mucosa that could be detected as none subject had normal histology [5,24]. Also considering limited sample size for each gastric lesion, we combined mild gastric lesions (SG and CAG) as reference for downstream analyses. The major outcome measures included the risk of GC and the progression of gastric lesions. We also examined the risk of advanced gastric lesions (IM or LGIN) as a secondary outcome.

2.6.1. Proteomic profiles of gastric lesions and GC

In the discovery stage, hierarchical clustering was conducted to visualize the characteristics of proteomic profiles across subjects of mild gastric lesions (SG or CAG), advanced gastric lesions (IM or LGIN), and GC. The analyses were restricted to proteins detected in more than 1/6 of all samples (Fig. 1a-1b). Hierarchical clustering was performed based on the Euclidian distance through complete method and proteins with similar trajectories were grouped (Fig. 1c). Gene ontology functional annotation analysis and pathway enrichment analysis were conducted for biological functions of each cluster by querying Metascape, combining the data resources of GO, KEGG, Uniprot and Drug bank (http://metascape.org).

2.6.2. Proteomic subtyping of precancerous gastric lesions

We explored the proteomic-based molecular subtypes of gastric lesions using the unsupervised non-negative matrix factorization consensus cluster package (NMF v.0.21.0) in R [29]. Coefficient of variation was calculated for each protein and was used to rank the proteins in a descending order. A preferred cluster result was selected by visual inspection, considering profiles of average silhouette width ranging between 2 and 8 clusters. A supervised random forest classifier was constructed based on discovery stage samples and applied to the validation stage to identify subject subgroups with the same proteomic signatures. The subtypes were not generated for GC considering our specific interest in exploring proteomic signatures that would benefit the prediction of the risk of gastric lesion progression prospectively. Logistic regression analysis was performed for the association of proteomic subtypes with the risk of gastric lesion progression. To distinguish differentially expressed proteins across proteomic subtypes, the quantitative iFOTs were transformed (log (1 +iFOT)) to conform a Gaussian or normal distribution and Student's t test was then used to identify subtype-specific proteins that met $P < 0.05$ and fold change between subtypes>1.

2.6.3. Association analyses of individual proteins

In the discovery stage, we examined the associations between individual protein expression and risk of advanced gastric lesions (IM or LGIN) and GC, with mild gastric lesions (SG or CAG) as the reference. The analyses were conducted for 2682 proteins that were identified in at least 1/2 samples of each comparison group. The ORs and 95% confidence intervals (CI) were calculated using logistic regression analyses, adjusting for age at diagnosis and sex [30]. *H. pylori* infection was not adjusted for as we lacked the information for GCs in the discovery stage. Missing values for these proteins were imputed using the 1/10 of minimum expression value of all proteins in each sample. Considering multiple comparisons, FDR-q<0.05 was set as the significance level for the discovery stage.

We sought to validate the associations of significant proteins in two validation stages using logistic regression models, adjusting for age and sex, with $P < 0.05$ and same direction of association with the discovery stage considered successful validation. We also examined the associations between key individual proteins and the risk of progression of gastric lesions, comparing progressed subjects (n = 19, which all progressed to IM, LGIN or HGIN, Table S2) with those non-progressed (n = 20), taking advantage of the prospective endoscopic follow-up of Linqu validation subjects. We further examined whether key individual proteins at baseline were associated with having IM or severer gastric lesions (IM, LGIN or HGIN) at endpoint. Again, logistic regression models were used, adjusting for age, sex, baseline pathology and *H. pylori* infection. As cohort follow-up completely relied on gastroendoscopy, the time at the end of follow-up may not represent the exact occurrence time of progression of gastric lesions. Therefore, logistic regression models were used instead of Cox regression models, as the latter requires clear time axis. For 21 cases of gastric lesions that had biopsies taken at follow-up endpoint, changes of protein expression level with the progression of gastric lesions were assessed using Wilcoxon rank-sum test.

For highlighted GC-associated proteins, we explored whether they would be associated with GC of intestinal or diffuse type differently and the overall survival of GC, based on discovery stage subjects. We also explored their associations with the proteomic subtypes of gastric lesions and GC, compared with the proteomic-defined mild gastric lesions. Logistic regression models were used, adjusting for age and sex.

For highlighted proteins in our study, we examined the differences in mRNA expression between 410 GC tissues and 35 non-tumor tissues based on The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/repository). Analyses were conducted using logistic regression adjusting for age and sex.

2.6.4. Construction of prediction models for the progression of gastric lesions

A risk score model was constructed by summing the weighted expression of key individual proteins validated to be associated with early GC and progression of gastric lesions. The weighted expression of each protein was calculated as the standardized expression of each protein (iFOT divided by the standard deviation of iFOTs for this protein) multiplied by the regression coefficient from logistic regression, using the below function:

$$\text{Riskscore} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_n X_n$$

We calculated a risk score for each subject of gastric lesions in the validation stage and tested how risk scores changed the risk of gastric lesion progression. The OR (95% CI) for standard deviation of risk score was calculated using logistic regression adjusted for age, sex, baseline histopathology and *H. pylori* infection.

Efforts were also made to construct prediction models using random forest classifier for progression of gastric lesions, integrating the risk score of key proteins and proteomic subtypes of gastric lesions, with other GC risk factors. Based on 39 cases of gastric lesions with prospective follow-up in Linqu validation set, we constructed a machine learning prediction model and estimated the performance by leave-one-out cross-validation method. The receiver operating characteristic curve was plotted to assess the predictive value. Delong's test was used to compare the performance of different prediction models. To facilitate the translation of findings, an online calculator was developed for predicting the risk of gastric lesion progression.

2.7. IHC staining and evaluation

IHC methodology was developed for detecting expression of APOA1BP, DDT, HPX, PGC in formalin-fixed, paraffin-embedded (FFPE) tissues and the assays were conducted for 65 subjects (12 SGs, 6 CAGs, 14 IMs, 11 LGINs, 9 early GCs, 13 invasive GCs). FFPE sections (4 μm) were deparaffinized in xylene and then hydrated in graded alcohol. After antigen retrieval, the specimens were blocked by 3% H_2O_2 for 1h. Samples were incubated with primary antibodies (anti-APOA1BP, ab199130, RRID: AB_2893269, 1:15; anti-DDT, ab115785, RRID: AB_10933538, 1:500; anti-HPX, ab124935, RRID: AB_10975463, 1:200; anti-PGC, ab180709, RRID: AB_2893268, 1:50; all from Abcam) overnight at 4 ℃, rinsed in PBS, then detected by secondary antibody (anti-rabbit IgG, ab6721, RRID_95547, 1:500, Abcam) for 1 h at room temperature, rinsed in PBS, and finally stained with DAB for 5 min. The positive percentage of staining area was evaluated by Image J based on three random captured visual fields by technicians blinded for sample groups. Brown stains were identified as specific antigen-antibody bindings and normal stromal cells were excluded from the analysis.

2.8. Role of funders

Funders had no role in study design, data collection, data analyses, interpretation or writing of report.

3. Results

3.1. Proteomic profiling of precancerous gastric lesions and GC

We identified a total of 15158 gene products at 1% FDR during the discovery stage and 9119 were deemed with high reliability (Fig. 1a). Assays of QC samples (293T cell) supported the stability of proteomic profiling (Fig.S1a). All samples showed good consistency in the quantification of tissue proteome (all Spearman's correlation $r > 0.80$, Fig. S1b).

3.2. Protein expression patterns with progression of gastric lesions

Unsupervised clustering of proteomic data was conducted for three pre-defined subject groups: mild gastric lesions (SG/CAG), advanced gastric lesions (IM/LGIN) and GC, which revealed distinct profiles from gastric lesions to GC, while the overall differences between mild and advanced gastric lesions appeared not quite striking (Fig. 1b). We then explored detailed protein expression patterns along the cascade of gastric lesion progression and extracted 6 protein clusters that possessed similar trajectories (Fig. 1c), which demonstrated dynamic changes in proteomic profiles through mild gastric lesions, advanced gastric lesions to GC.

Among aggregated clusters, cluster-1 proteins were highly expressed in advanced gastric lesions and particularly enriched in the pathways of biological oxidations and cellular amino acid metabolic process. Intestine-specific proteins (MUC2, FABP1, MYO7B, ANXA13 and CDH17) annotated in the Human Protein Atlas [31] were enriched in cluster-1. Cluster-2 proteins were enriched in the digestion and metabolism of carbohydrates pathways and exhibited decreased expression from subjects with mild to advanced gastric

lesions, and then a sharp decrease to GC. Known stomach-specific proteins [31], such as PGC, MUC5AC, MUC1, TFF1, TFF2, VSIG1, VSIG2, CTSE, ANXA10 were aggregated in cluster-2. In contrast, cluster-3 proteins displayed increased expression with severity of gastric lesions and were enriched in the pathways of JAK-STAT signaling after interleukin-12 stimulation (e.g. CA1 and SP100) and oxygen transport (e.g. SLC4A1 and PSME2). Proteins in cluster-4 were enriched in vesicle-mediated transport (e.g. HPX and IGF2R) and cytokine signaling in immune system pathways (e.g. PSMC5 and STAT6) and presented stable changes in gastric lesions but a sharp increase in GC. Proteins in cluster-5 were enriched in the pathways of nucleobase-containing small molecule metabolic process (e.g. APOA1BP and SLC44A2) and carboxylic acid catabolic process (e.g. GLUL and ALDH6A1). Proteins in cluster-6 were enriched in pathways of smooth muscle contraction (e.g. COX5A and CALM1) and negative regulation of macrophage migration (e.g. DDT and MIF). Both cluster-5 and cluster-6 appeared undulated from mild to advanced gastric lesions and then to GC (Fig. 1c).

3.3. Proteomic-based molecular subtypes of gastric lesions

We explored molecular similarity and heterogeneity of gastric lesions (SG, CAG, IM, or LGIN) beyond cellular morphology level. Based on top 100 most-variant proteins detected in more than 3/4 subjects of gastric lesions, we derived 4 molecular subtypes (S1-S4) of gastric lesions with a best average silhouette-width (0.97) through NMF algorithm (Fig.S2). A significant correlation was found between molecular subtypes and severity of pathological diagnosis (Spearman's correlation r = 0.42, $P = 5.62\times10^{-6}$), with subtype-S1 representing proteomic-defined mild gastric lesion and S4 representing the most severe gastric lesion. The distribution of proteomic subtypes appeared independent from *H.pylori* infection. Subtype-specific proteins (Student's t test P<0.05 and fold change between subtypes>1) are shown in the heatmap (Fig. 2a). Compared with subtype-S1, proteins highly expressed in other subtypes were enriched in apoptotic process.

We constructed a machine learning classifier for four proteomic subtypes based on discovery stage subjects and applied the classifier to validation set subjects. Logistic regression analyses based on prospective follow-up of Linqu validation subjects found that subtype-S4 had the highest risk of gastric lesion progression (OR = 19.29, 95% CI: 1.82-204.59 vs. subtype-S1).

3.4. Key individual proteins associated with gastric lesion progression and early GC

We explored key individual proteins for clues on potential convenient biomarkers. We did not find proteins associated with the risk of advanced gastric lesions (IM/LGIN) at FDR-q<0.05 (logistic regression analysis). In contrast, 1201 proteins were significantly associated with the risk of invasive GC in the discovery stage (FDR-q<0.05, logistic regression analysis) and validated 217 for the risk of early GC in Linqu validation set ($P < 0.05$, logistic regression analysis), including 104 positively-associated and 113 inversely-associated proteins (Fig. 2b, Table 1 and Table S3). The analyses restricting to HGINs (specified as 'very early' GC here) compared with mild gastric lesions did not alter the findings.

We focused on the 217 proteins for further analyses. The top pathways enriched for these proteins are shown in Fig.S3. We explored the associations of 217 individual proteins with the odds of proteomic subtypes of gastric lesions and GC. Due to modest sample size, subtypes of S2, S3, and S4 were combined as proteomic-defined advanced gastric lesions. Compared with subtype-S1, 25 out of 217 proteins were significantly associated with proteomic-defined advanced gastric lesions (subtypes S2-S4) and GC in both the discovery (FDR-q<0.05, logistic regression analysis) and two validation sets ($P < 0.05$, logistic regression analysis) (Table S4).

Leveraging prospective follow-up of subjects in the Linqu validation set, we examined the associations between 217 individual proteins as detailed above and risk of gastric lesion progression. Four proteins, namely APOA1BP, PGC, HPX and DDT, were further significantly associated with the risk of gastric lesion progression and the risk of having IM or more severe gastric lesions at the follow-up endpoint ($P < 0.05$, logistic regression analysis). HPX was positively associated with early GC and gastric lesion progression, while APOA1BP, PGC and DDT were inversely associated (Fig. 2b-2c and Table 1). These four highlighted proteins were aggregated in the cluster-2 (PGC), -4 (HPX), -5 (APOA1BP) and -6 (DDT) respectively (Fig. 1c). We also examined the dynamic changes in protein expression during follow-up and found significantly decreased expression of APOA1BP and PGC only among progressed subjects during follow-up ($P < 0.05$, Wilcoxon rank-sum test). Due to limited sample size, we didn't find significant differences of HPX and DDT between baseline and follow-up endpoint, but still there was a trend towards increased expression of HPX and decreased expression of DDT at the endpoint of progressed subjects, consistent with that observed in association analyses (Fig. 2d).

The associations of APOA1BP, PGC, HPX and DDT with early GC were further validated in the independent Beijing validation set (Table 1). Pooling two validation sets together to increase the sample size, DDT, PGC and APOA1BP showed significantly inverse associations with early GC while HPX showed significantly positive associations (Fig. 3a and Table 1).

To validate the associations of four key molecules with GC at mRNA level, we examined mRNA expression of these proteins based on TCGA dataset. Consistent with the associations for proteins, we found increased mRNA expression of HPX and decreased expression of APOA1BP, DDT and PGC in GCs compared with adjacent normal tissues, although the association was not statistically significant for APOA1BP (Table S5).

In a secondary analysis, we found no heterogeneity in the associations of four highlighted proteins with intestinal- or diffuse-type GC based on discovery stage subjects (Table S6). These four proteins were not significantly associated with the overall survival of GC (Table S7).

3.5. Construction of risk score and prediction models for gastric lesion progression

Combining four key proteins which stood out in the prospective analyses, a risk score model was derived: Risk score = $-1.485 \times APOA1BP - 1.231 \times PGC + 1.686 \times HPX - 0.565 \times DDT$. The risk score was independently associated with risk of gastric lesion progression based on Linqu validation cohort, with an OR of 4.09 (95% CI: 1.48-11.27, logistic regression analysis) per one-standard-deviation of risk score increase (Fig. 3a).

Integrating proteomic signatures, we constructed machine learning prediction models for progression of gastric lesions. Compared with the model including age, sex, *H.pylori* infection and baseline pathology (model-1, areas under the curve (AUC)=0.56, 95%CI: 0.36-0.76), the prediction model integrating the risk score of four proteins performed better (model-2, AUC=0.79, 95%CI: 0.65-0.93, Delong's test P=0.02 for model-2 vs. model-1). Additionally, integrating proteomic-defined molecular subtypes of gastric lesions, the model had further significantly improved performance in predicting progression of gastric lesions (model-3, AUC=0.88, 95%CI: 0.78-0.99, Delong's test P=0.002 for model-3 vs. model-1, Delong's test P = 0.04 for model-3 vs. model-2) (Fig. 3b). A user-friendly online calculator was developed to preliminarily calculate the progression probability of gastric lesions (https://www.aboutproteomics.com/predictprogression/).

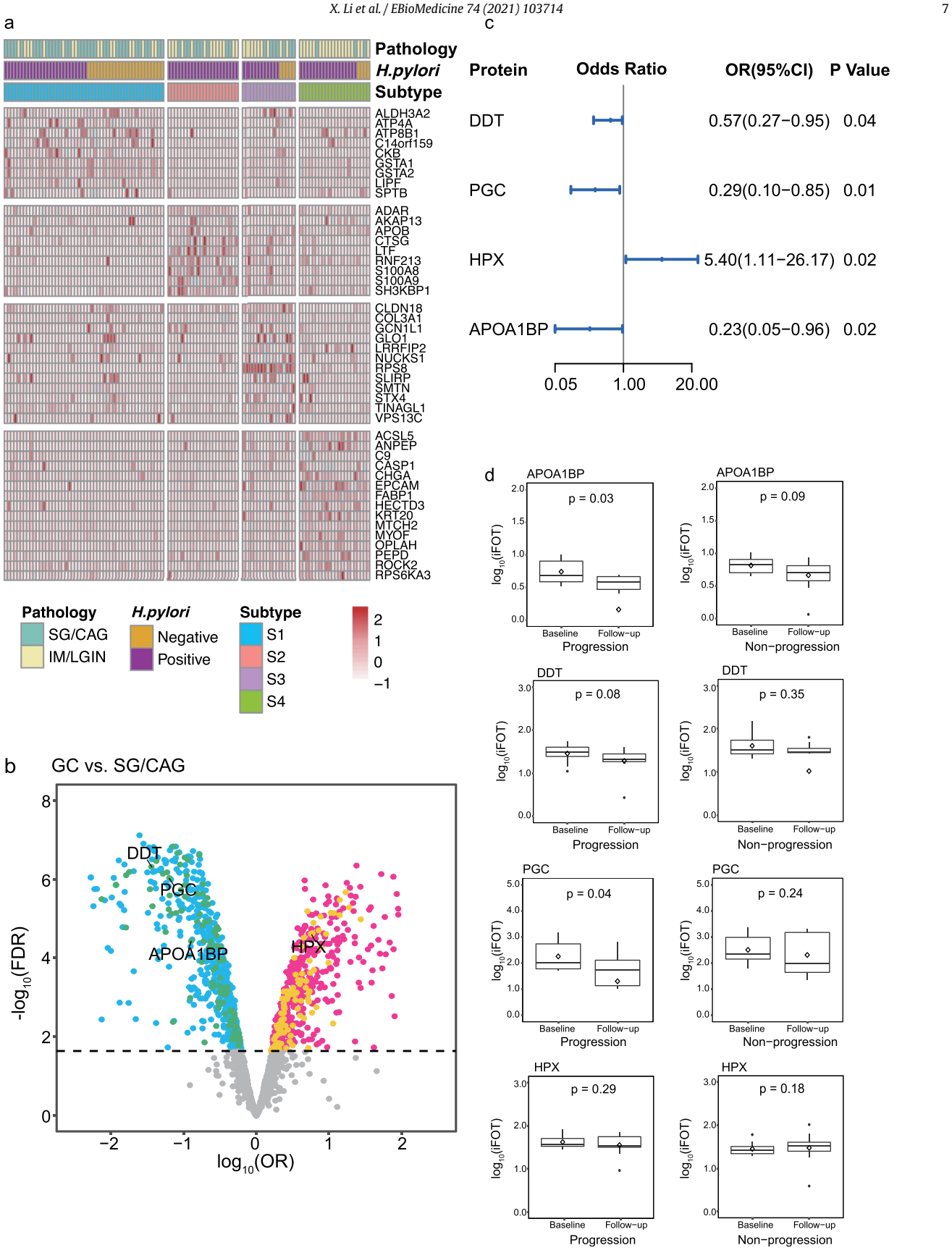

Fig. 2. Proteomic subtyping of precancerous gastric lesions and proteomic signatures associated with GC and gastric lesion progression. **a.** Heatmap of subtypes identified through unsupervised non-negative matrix factorization consensus clustering. Subtype-specific proteins with P<0.05 (Student's t test) and fold change between subtypes >1 are shown. **b.** Proteins significantly associated with GC. Proteins significantly associated with GC in the discovery stage (FDR-q<0.05, logistic regression analysis) were presented in pink (positive) and blue (inverse association) dots, and proteins validated for early GC were presented in yellow (positive) and green (inverse association) dots. The four proteins significantly associated with gastric lesion progression in prospective follow-up were labeled. **c.** Associations of four proteins with the risk of gastric lesion progression based on prospective follow-up. **d.** The expression of four proteins at baseline and follow-up endpoint for progressed (n = 10) and non-progressed (n = 11) subjects during follow-up. CAG, chronic atrophic gastritis; CI, confidence interval; *H. pylori*, *Helicobacter pylori*; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; OR, odds ratio; SG, superficial gastritis.

Table 1
Proteins significantly associated with risk of early gastric cancer and gastric lesion progression during the follow-up[a]

	Discovery set					Linqu validation set				Follow-up of Linqu validation set[b]				Beijing validation set				Combined validation sets			
	IM/LGIN (n = 59) vs. SG/CAG (n = 52)			GC (n = 58) vs. SG/CAG (n = 52)		IM/LGIN (n = 21) vs. SG/CAG (n = 18)		Early GC (n = 17) vs. SG/CAG (n = 18)		Progression (Yes, n = 19 vs. No, n = 20)		IM/LGIN/HGIN atendpoint (Yes, n = 27 vs. No, n = 12)[c]		IM/LGIN (n = 45) vs. SG/CAG (n = 24)		Early GC (n = 30) vs. SG/CAG (n = 24)		IM/LGIN (n = 66) vs. SG/CAG (n = 42)		Early GC (n = 47) vs. SG/CAG (n = 42)	
	OR(95%CI)	P	FDR	OR(95%CI)	P	FDR	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P	OR(95%CI)	P
DDT	1.16 (0.76-1.75)	0.49	0.81	0.04 (0.01-0.13)	4.89×10^{-7}	3.37×10^{-5}	0.72 (0.47-1.10)	0.10	0.39 (0.15-0.99)	0.048	0.57 (0.34-0.95)	0.04	0.64 (0.41-0.98)	0.04	0.74 (0.44-1.24)	0.13	0.50 (0.26-0.96)	0.04	0.71 (0.54-0.94)	0.02	0.52 (0.33-0.83)	0.01		
PGC	0.33 (0.19-0.57)	7.95×10^{-5}	0.12	0.06 (0.02-0.19)	9.34×10^{-7}	3.74×10^{-5}	0.23 (0.09-0.59)	0.005	0.20 (0.07-0.60)	0.008	0.29 (0.12-0.71)	0.01	0.46 (0.25-0.86)	0.02	0.97 (0.51-1.85)	0.46	0.001 (2.85×10^{-5}-0.04)	0.001	0.60 (0.40-0.90)	0.02	0.09 (0.03-0.31)	0.001		
HPX	1.45 (0.69-3.05)	0.33	0.72	5.80 (2.56-13.14)	2.52×10^{-5}	2.35×10^{-4}	0.61 (0.18-2.04)	0.25	4.21 (1.05-16.84)	0.04	5.40 (1.43-20.30)	0.02	6.49 (1.20-35.09)	0.03	1.79 (0.73-4.40)	0.10	2.23 (1.11-4.48)	0.03	1.38 (0.76-2.49)	0.19	2.61 (1.42-4.80)	0.005		
APOA1BP	0.71 (0.45-1.11)	0.13	0.55	0.13 (0.05-0.34)	3.74×10^{-5}	3.09×10^{-4}	0.37 (0.13-1.01)	0.05	0.18 (0.04-0.76)	0.02	0.23 (0.07-0.76)	0.02	0.31 (0.10-0.92)	0.04	0.20 (0.05-0.78)	0.01	0.76 (0.46-1.25)	0.18	0.30 (0.16-0.57)	0.001	0.54 (0.33-0.91)	0.02		

CAG, chronic atrophic gastritis; GC, gastric cancer; HGIN, high-grade intraepithelial neoplasia; *H. pylori*, *Helicobacter pylori*; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; OR, odds ratio; SG, superficial gastritis.

[a] Proteins with FDR<0.05 associated with GC in the discovery stage and P<0.05 associated with early GC in validation stage and $P < 0.05$ associated with gastric lesion progression in follow-up are shown here. Analyses were conducted using unconditional logistic regression adjusting for age and sex.

[b] *H. pylori* infection and baseline histopathological diagnosis were additionally adjusted for.

[c] Subjects with gastric lesions progressing to IM, LGIN, HGIN or having these lesions at the follow-up endpoint were compared with other subjects.

3.6. Validation of four key proteins by IHC assays

To help the translation of findings to clinical and preventive settings, we developed IHC assays for four key proteins based on 65 FFPE tissues. All four proteins are predominantly stained in cytoplasm. APOA1BP, DDT and PGC showed significantly decreased expression in HGIN and invasive GC, and HPX showed significantly increased expression in HGIN and invasive GC, consistent with findings in proteomic assays (Fig. 3c).

4. Discussion

To our knowledge, this is the first study that comprehensively explored dynamic changes of proteome and protein signatures in the evolution of gastric lesions and early GC. We revealed heterogeneity of gastric lesions in proteomic features and defined four proteomic subtypes associated with different risk of gastric lesion progression. PGC, HPX, APOA1BP and DDT were significantly associated with the risk of early GC and HGIN (specified here as 'very early' GC) and the progression of gastric lesions, based on prospective follow-up of population-based subjects.

In high-risk areas for GC such as Linqu, majority of GCs are of intestinal type, which is believed to be preceded by a prolonged precancerous process [32]. Previous studies have assessed molecular subtypes of GC based on -omics approaches [6–10]. For example, TCGA research network reported four clinically relevant molecular subtypes of GC [7]. However, these pioneering efforts did not include subjects with known precancerous gastric lesions. The molecular features of gastric lesions were undefined and risk of gastric lesion progression to GC needs to be clarified according to molecular subtypes, beyond cellular morphology level. No studies have examined molecular signatures along the cascade of evolution of precancerous gastric lesions and GC development, pursuing improving the ability to precision GC prevention and management. In prior studies of GC [8] and colorectal cancer [33], a considerable number of genes with DNA mutations did not have their gene products detected in proteome. Although efforts were made to construct mRNA profiles of individual cells in GC and several gastric lesions [34], mRNA transcript abundance may not reliably predict the differences of proteins [33,35,36]. As a plausible perspective to explore GC carcinogenesis, clarification of key protein signatures may provide clues to GC etiology, aiding the discovery of intervention targets. Interpretation of proteomic signatures for precancerous gastric lesions and their progression may help define novel biomarkers that could be used to uncover a population at markedly increased GC risk. A prospective study design is favorable for realizing this goal, which stands at a higher level for evidence inference than cross-sectional or case-control design commonly used previously.

We observed distinct proteomic profiles between subjects with precancerous gastric lesions and GC. Heterogeneity in molecular features was further revealed for gastric lesions based on the four newly derived proteomic subtypes. We validated a number of proteins associated with proteomic-defined advanced gastric lesions (subtypes S2-S4) and GC risk, compared with subtype-S1 as a signature for mild gastric lesions. None of these proteins were associated with histopathology-determined advanced gastric lesions (IM or LGIN) in the discovery stage at FDR-q<0.05. Altogether, proteomic-based subtypes may open a new avenue for assessing the severity of gastric lesions. An approach integrating molecular pathological assessment may provide new insights into gastric lesions, beyond cellular morphology pathological level. The model integrating four individual proteins and proteomic-defined molecular subtypes of gastric lesions had significantly improved performance in predicting progression of gastric lesions, corroborating the potential translational significance of these subtypes.

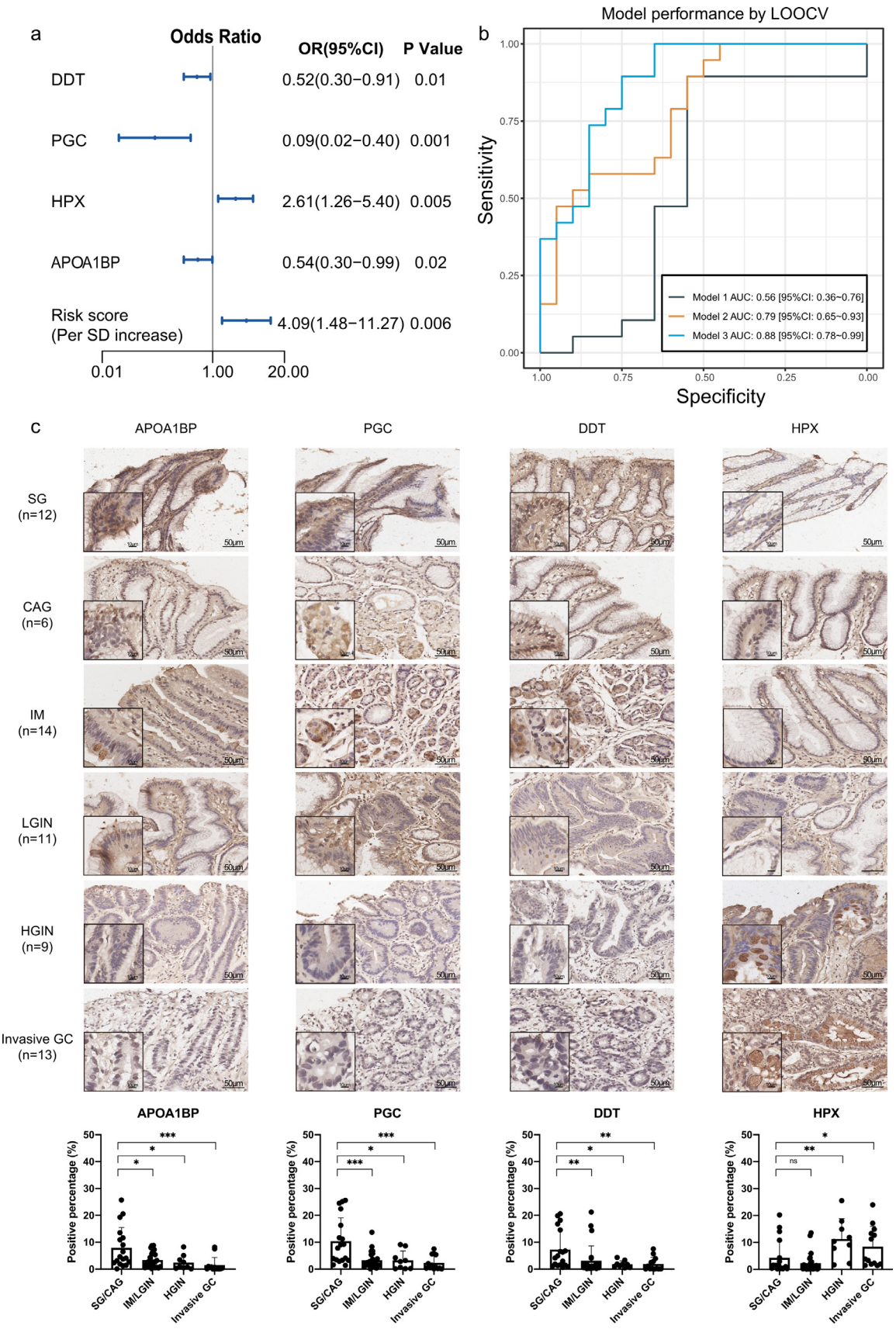

Fig. 3. Validation of proteomic signatures and risk prediction models for gastric lesion progression. **a** Associations of four proteins with early GC in combined validation sets and association of risk score with progression of gastric lesions in prospective follow-up subjects. **b.** Receiver operating characteristic curve of random forest classifier prediction model for progression of gastric lesions during the follow-up. Performance of the models was estimated by leave-one-out cross-validation method. Model-1 includes age, sex, *H.pylori* infection and baseline pathology. Model-2 includes variables in model-1 and risk score of four proteins. Model-3 includes all variables in model-2 and molecular subtypes. **c.** Immunohistochemistry staining of four proteins in formalin-fixed, paraffin-embedded tissues (×40 magnification). All four proteins are predominantly stained in cytoplasm, consistent

In our study, subtype-S4 showed high expression of markers of the epithelial lineage (EPCAM, KRT20, FABP1), while S2 displayed high expression of markers of monocytes (S100A8, S100A9), and S3 showed high expression of COL3A1, a marker of fibroblasts. All biopsies were taken according to standardized endoscopic protocols, were similar in size, and contained mainly mucosa with small part of underlying submucosa. The large number of samples in our study compensates for a potential bias based in minor differences in tissue composition. The identification of proteomic subtypes from unbiased analysis using bulk proteomics suggest that non-cancerous gastric lesions may have generated their own "microenvironment", namely S2 with monocyte, and S3 with fibroblast, while S4 with epithelial cells, providing insights for our understanding of gastric lesions. This indicates that along with the changes in epithelial cells, ongoing alterations of the tissue microenvironment, including macrophages and fibroblasts, may also be involved in GC carcinogenesis.

We validated four proteins including APOA1BP, PGC, HPX and DDT associated with the risk of early GC and progression of gastric lesions based on prospective follow-up. Of them, pepsinogen C (PGC), the precursor of pepsin C with known function of digestion, has demonstrated potential importance for diagnosis and prognosis prediction of GC previously [37,38]. Several studies have reported declined PGC expression in GC compared with controls [37,39,40] and a sequential decrease in SG and CAG [40]. A case-control study by Repetto et al initially used two-dimensional difference gel electrophoresis and then used LC-MS/MS to identify protein signatures [41]. Although with a relatively limited sample size (n = 60) and less coverage of protein profiles, PGC was also found down-regulated in atrophic gastritis and GC [41]. In a prospective study, Ning et al reported PGC-MG7 panel possibly useful for detecting high-risk populations for GC [42], but this biomarker combination still requires validation in other studies.

Other three proteins have also indicated potential importance previously. APOA1BP is secreted into gastric fluid and interacts with apolipoprotein A-I. A proteomic study also reported downregulated APOA1BP in GCs [43]. We further found decreased PGC and APOA1BP expression in advanced gastric lesions and GC compared with mild gastric lesions, and lowered expression among individuals with progression of gastric lesions. Prior evidence on HPX and DDT have been sparse for GC. Hemopexin (HPX) was reported to defend against oxidative stress and related inflammatory disorders [44]. DDT (D-dopachrome tautomerase, also known as MIF-2) is a member of MIF family and loss of MIF in mice was shown to promote tumor development [45]. In our study, although the expression of these two proteins did not alter significantly in IM or LGIN, the association of elevated HPX expression and lowered DDT expression with early GC, even with HGIN, appeared robust based on our analysis of the discovery stage and two validation stages, in both case-control and prospective studies. Whether these proteins may serve as druggable targets remains be elucidated. Although three of the highlighted proteins, including APOA1BP, PGC and DDT, are negative markers, they had reasonable expression abundance in advanced gastric lesions and early GC. In addition, the study was designed to focus on precancerous gastric lesions and early GC, as part of our ongoing efforts to promote high-risk population identification and early GC detection. A risk score combining these four proteins, along with proteomic subtypes of gastric lesions, were both associated with the risk of gastric lesion progression. In addition, IHC methods were developed to display the tissue and cellular localization, which replicated the associations for four proteins, facilitating the translation of findings in GC prevention and clinical settings.

Both tissue and serum (or plasma) may serve as appropriate biospecimens for potential biomarkers. Tissues are very commonly used for biomarker detection utilizing surgically resected samples or biopsies. For example, the detection of HER2 or PD-L1 by IHC or in situ hybridization is recommended for deciding treatment options of metastatic GC according to the new NCCN guideline [46]. Based on this guideline, the use of NGS liquid biopsy is necessary only when tissues cannot be collected [46]. While the screening for serum biomarkers has been long pursued, available serum biomarkers such as CA19-9 and CEA are far from sensitive and specific for GC prevention and management [47,48]. CA19-9 and CEA are even undetectable in gastric tissues in our study. In addition, markers for risk of precancerous lesions and their progression may not necessarily be found in the serum. The tissue proteomic signatures that we identified were significant to predict the risk of progression of gastric lesions and GC occurrence. The application of tissue proteomic signatures may potentially help healthy providers decide for appropriate prevention and management strategies. MS can provide the solution in proteomics labs and well-developed IHC assays may extend the use into most hospitals. Revealing proteomic-based molecular features of gastric lesions beyond cellular pathology would contribute to defining the 'real' high-risk population for precision primary GC prevention and identifying early GC for efficient secondary prevention.

Strengths of our study included proteomic profiling with relatively large sample size, two-stage validation with prospective cohort study design, validation at mRNA level (TCGA) and based on IHC staining. Our study offered compelling leads for these proteins as early biomarkers for GC detection. We acknowledge several limitations. First, although we made the first attempt to conduct a prospective study for gastric lesions, we had a modest sample size of subjects with longitudinal follow-up and haven't been able to thoroughly follow the subjects with gastric lesions of the discovery set and Beijing validation set by endoscopy. Second, we did not include any subjects without gastric lesions. However, population-based endoscopic screening in Linqu in this and previous studies can barely find any individual with completely normal gastric histology [5,24]. Indeed, the claimed 'normal' controls in public datasets basically are 'non-cancers' lacking detailed pathological examinations. Despite so, future studies are warranted to explore proteomic profiles in normal stomach. Third, the discovery stage only included cases of invasive GC, mostly at late clinical stages. However, all proteomic assays were conducted following the same standard procedures; we deliberately pursued validation of results only involving early GC and were able to replicate the findings using two validation sets. Fourth, our study cannot directly answer mechanisms underlying the associations. The visualized protein clusters and their enriched pathways were only exploratory in nature. Although we have confirmed the direction of associations for highlighted proteins at mRNA level in TCGA, we lack direct mRNA expression data in our study.

While the current GC prevention program in China relies fully on gastroscopy screening following a 'one-fits-all' strategy, our study may have translational implications facilitating the implementation of precision GC prevention. Particularly high-risk populations for GC as uncovered by proteomic signatures may benefit from targeted primary interventions to prevent GC occurrence and frequent screening for GC identification at a very early stage, advancing the primary and secondary GC prevention. Multi-center, larger-scale prospective studies would be warranted before regulatory approval of wide application of key protein signatures.

with the Human Protein Atlas report (https://www.proteinatlas.org/ENSG00000110169-HPX/pathology/stomach+cancer#Location for HPX, ENSG00000096088 for PGC, ENSG00000099977 for DDT, and ENSG00000163382 for APOA1BP). Wilcoxon rank-sum test, ns=non-significant, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. AUC, area under the curve; CAG, chronic atrophic gastritis; CI, confidence interval; GC, gastric cancer; HGIN, high-grade intraepithelial neoplasia; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; LOOCV, leave-one-out cross-validation; OR, odds ratio; SD, standard deviation; SG, superficial gastritis.

Contributors

WQL, JQ and KFP had verified the underlying data and had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. WQL and JQ supervised and designed the research; XL, NRZ, ZCL, HF, JD, XTN, and XW adapted algorithms and software for data analyses; XL, ZXL, HF, TZ, YZ, GK, SH, WDL, LFZ, JMX and LHW contributed to subject recruitment and sample collection; ZWL and XZW completed histological diagnoses; JYZ and WHW performed *H. pylori* antibody assays; MWL and KL carried out sample preparation and mass spectrometry analyses; XL wrote the draft of the manuscript; WQL, JQ, KFP, MG, YW and WCY revised the manuscript. All authors read and approved the submitted version.

Declaration of Competing Interest

XL, NRZ, YW and JQ have a pending patent entitled "Proteomic subtyping of precancerous gastric lesions, biomarkers associated with gastric cancer and gastric lesion progression and prediction models for gastric lesion progression" (CNIPA patent application number: 202010958039X, filed September 11, 2020). The other authors have declared that no conflict of interest exists.

Acknowledgements

We are indebted and thankful to all participants for their valuable contributions. This work was supported by Beijing Talents foundation (2018000021223ZK01), Michigan Medicine-PKUHSC Joint Institute for Translational and Clinical Research (BMU2020JI004), Capital's Funds for Health Improvement and Research (CFH 2020-2-1026), National Natural Science Foundation of China (NSFC-DFG, 81861138041), National Key Research and Development Program of China (2018YFA0507503) and Beijing Municipal Administration of Hospitals' Ascent Plan (DFL20181102). The mRNA level analysis is based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We thank the contribution of the appropriate specimen donors and research groups.

Data sharing statement

The MS proteomics data are deposited to the ProteomeXchange Consortium via the iProX partner repository (access No. IPX0003438002, https://www.iprox.cn/page/subproject.html?id=IPX0003438002).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103714.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
[2] Fan X QX, Zhang Y, Li Z, Zhou T, Zhang J, You W, Li W, Pan K. Screening for gastric cancer in China: Advances, challenges and visions. Chin J Cancer Res 2021;33(2):168–80.
[3] You WC, Blot WJ, Chang YS, Ershow AG, Yang ZT, An Q, et al. Diet and high risk of stomach cancer in Shandong, China. Cancer Res 1988;48(12):3518–23.
[4] You WC, Li JY, Blot WJ, Chang YS, Jin ML, Gail MH, et al. Evolution of precancerous lesions in a rural Chinese population at high risk of gastric cancer. Int J Cancer 1999;83(5):615–9.
[5] Li WQ, Zhang JY, Ma JL, Li ZX, Zhang L, Zhang Y, et al. Effects of Helicobacter pylori treatment and vitamin and garlic supplementation on gastric cancer incidence and mortality: follow-up of a randomized intervention trial. BMJ 2019;366:l5016.
[6] Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015;21(5):449–56.
[7] Cancer genome atlas research N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513(7517):202–9.
[8] Ge S, Xia X, Ding C, Zhen B, Zhou Q, Feng J, et al. A proteomic landscape of diffuse-type gastric cancer. Nat Commun 2018;9(1):1012.
[9] Mun DG, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell 2019;35(1):111-24 e10.
[10] Ni X, Tan Z, Ding C, Zhang C, Song L, Yang S, et al. A region-resolved mucosa proteome of the human stomach. Nat Commun 2019;10(1):39.
[11] Kojima T, Yoshikawa K, Saga S, Yamada T, Kure S, Matsui T, et al. Detection of elevated proteins in peritoneal dissemination of gastric cancer by analyzing mass spectra data of serum proteins. J Surg Res 2009;155(1):13–7.
[12] Liu C, Pan C, Liang Y. Screening and identification of serum proteomic biomarkers for gastric adenocarcinoma. Exper Ther Med 2012;3(6):1005–9.
[13] Mohri Y, Mohri T, Wei W, Qi YJ, Martin A, Miki C, et al. Identification of macrophage migration inhibitory factor and human neutrophil peptides 1-3 as potential biomarkers for gastric cancer. Br J Cancer 2009;101(2):295–302.
[14] Qiu FM, Yu JK, Chen YD, Jin QF, Sui MH, Huang J. Mining novel biomarkers for prognosis of gastric cancer with serum proteomics. J Exper Clin Cancer Res 2009;28:126.
[15] Song D, Yue L, Li H, Zhang J, Yan Z, Fan Y, et al. Diagnostic and prognostic role of serum protein peak at 6449 m/z in gastric adenocarcinoma based on mass spectrometry. Br J Cancer 2016;114(8):929–38.
[16] Wu W, Juan WC, Liang CR, Yeoh KG, So J, S100A9 Chung MC. GIF and AAT as potential combinatorial biomarkers in gastric cancer diagnosis and prognosis. Proteomics Clin Appl 2012;6(3-4):152–62.
[17] Wu W, Yong WW, MC Chung. A simple biomarker scoring matrix for early gastric cancer detection. Proteomics 2016;16(22):2921–30.
[18] Yang L, Wang J, Li J, Zhang H, Guo S, Yan M, et al. Identification of serum biomarkers for gastric cancer diagnosis using a human proteome microarray. Mole Cell Proteomics 2016;15(2):614–23.
[19] Fernández-Coto DL, Gil J, Hernández A, Herrera-Goepfert R, Castro-Romero I, Hernández-Márquez E, et al. Quantitative proteomics reveals proteins involved in the progression from non-cancerous lesions to gastric cancer. J Proteomics 2018;186:15–27.
[20] Sousa JF, Ham AJ, Whitwell C, Nam KT, Lee HJ, Yang HK, et al. Proteomic profiling of paraffin-embedded samples identifies metaplasia-specific and early-stage gastric cancer biomarkers. Am J Pathol 2012;181(5):1560–72.
[21] Li P, Ma D, Zhu ST, Tang XD, ST Zhang. Serum peptide mapping in gastric precancerous lesion and cancer. J Digest Dis 2014;15(5):239–45.
[22] Wang FR, Wei YC, Han ZJ, He WT, Guan XY, Chen H, et al. Aberrant DNA-PKcs and ERGIC1 expression may be involved in initiation of gastric cancer. World J Gastroenterol 2017;23(33):6119–27.
[23] Zhang L, Blot WJ, You WC, Chang YS, Kneller RW, Jin ML, et al. Helicobacter pylori antibodies in relation to precancerous gastric lesions in a high-risk Chinese population. Cancer Epidemiol Biomarkers Prevent. 1996;5(8):627–30.
[24] You WC, Blot WJ, Li JY, Chang YS, Jin ML, Kneller R, et al. Precancerous gastric lesions in a population at high risk of stomach cancer. Cancer Res 1993;53(6):1317–21.
[25] Dixon MF, Genta RM, Yardley JH, Correa P. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. Am J Surg Pathol 1996;20(10):1161–81.
[26] Feng J, Ding C, Qiu N, Ni X, Zhan D, Liu W, et al. Firmiana: towards a one-stop proteomic cloud platform for data processing and analysis. Nat Biotechnol 2017;35(5):409–12.
[27] Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature 2011;473(7347):337–42.
[28] Zhang CC, Chen Y, Mao XF, Huang Y, Jung SY, Jain A, et al. A bioinformatic algorithm for analyzing cell signaling using temporal proteomic data. Proteomics 2017;17(22).
[29] Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. PNAS 2004;101(12):4164–9.
[30] Thrift AP, HB El-Serag. Burden of gastric cancer. Clin Gastroenterol Hepatol 2020;18(3):534–42.
[31] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347(6220):1260419.
[32] Correa P. Human gastric carcinogenesis: a multistep and multifactorial process–First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res 1992;52(24):6735–40.
[33] Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513(7518):382–7.
[34] Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep 2019;27(6):1934–1947.e5.
[35] Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534(7605):55–62.
[36] Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell 2016;166(3):755–65.
[37] Kon OL, Yip TT, Ho MF, Chan WH, Wong WK, Tan SY, et al. The distinctive gastric fluid proteome in gastric cancer reveals a multi-biomarker diagnostic profile. BMC Med Genomics 2008;1:54.

[38] Shen S, Jiang J, Yuan Y. Pepsinogen C expression, regulation and its relationship with cancer. Cancer Cell Int 2017;17:57.
[39] Melle C, Ernst G, Schimmel B, Bleul A, Kaufmann R, Hommann M, et al. Characterization of pepsinogen C as a potential biomarker for gastric cancer using a histoproteomic approach. J Proteome Res 2005;4(5):1799–804.
[40] Ning PF, Liu HJ, Yuan Y. Dynamic expression of pepsinogen C in gastric cancer, precancerous lesions and Helicobacter pylori associated gastric diseases. World J Gastroenterol 2005;11(17):2545–8.
[41] Repetto O, De Re V, Giuffrida P, Lenti MV, Magris R, Venerito M, et al. Proteomics signature of autoimmune atrophic gastritis: towards a link with gastric cancer. Gastric Cancer 2021;24(3):666–79.
[42] Ning P, Sun L, Dong N, Yuan Y. PGC-MG7 combination could be used as a follow-up panel for monitoring dynamical progression of gastric precancerous diseases. Chin J Cancer Res 2020;32(1):89–95.
[43] Nishigaki R, Osaki M, Hiratsuka M, Toda T, Murakami K, Jeang KT, et al. Proteomic identification of differentially-expressed genes in human gastric carcinomas. Proteomics 2005;5(12):3205–13.
[44] Mehta NU, ST Reddy. Role of hemoglobin/heme scavenger protein hemopexin in atherosclerosis and inflammatory diseases. Curr Opin Lipidol 2015;26(5):384–7.
[45] Nemajerova A, Mena P, Fingerle-Rowson G, Moll UM, Petrenko O. Impaired DNA damage checkpoint response in MIF-deficient mice. EMBO J 2007;26(4):987–97.
[46] Ajani JA, D'Amico TA, Almhanna K, Bentrem DJ, Chao J, Das P, et al. Gastric cancer, version 3.2016, NCCN clinical practice guidelines in oncology. J Natl Comprehen Cancer Network 2016;14(10):1286–312.
[47] Feng F, Tian Y, Xu G, Liu Z, Liu S, Zheng G, et al. Diagnostic and prognostic value of CEA, CA19-9, AFP and CA125 for early gastric cancer. BMC Cancer 2017;17(1):737.
[48] Shimada H, Noie T, Ohashi M, Oba K, Takahashi Y. Clinical significance of serum tumor markers for gastric cancer: a systematic review of literature by the task force of the Japanese gastric cancer association. *Gastric Cancer* 2014;17(1):26–33.