Cancer Biol Med 2025. doi: 10.20892/j.issn.2095-3941.2025.0077

ORIGINAL ARTICLE

Proteomic profiling and scRNA sequencing identify signatures associated with *Helicobacter pylori* infection and risk of developing gastric cancer

Yu Jin[1]*, Xue Li[2]*, Bingyao Cai[3], Lanxin Yang[1], Wenjing Zhao[1], Hengmin Xu[1], Yang Zhang[4], Zongchao Liu[1], Kaifeng Pan[1], Wenqing Li[1]

[1]State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Peking University Cancer Hospital & Institute, Beijing 100142, China; [2]Department of Cancer Prevention, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou 310022, China; [3]School of Public Health, Peking University, Beijing 100191, China; [4]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing 100142, China

ABSTRACT

Objective: The key molecular events signifying the *Helicobacter pylori*-induced gastric carcinogenesis process are largely unknown.

Methods: Bulk tissue-proteomics profiling were leveraged across multi-stage gastric lesions from Linqu (n = 166) and Beijing sets (n = 99) and single-cell transcriptomic profiling (n = 18) to decipher key molecular signatures of *H. pylori*-related gastric lesion progression and gastric cancer (GC) development. The association of key proteins association with gastric lesion progression and GC development were prospectively studied building on follow-up of the Linqu set and UK Biobank (n = 48,529).

Results: Concordant proteomics signatures associated with *H. pylori* infection and gastric carcinogenesis (ρ = 0.784, correlation $P = 1.80 \times 10^{-36}$) were identified. RNA expression of genes encoding 13 up- and 15 down-regulated key proteins displayed trending alterations in the transition from normal gastric epithelium to intestinal metaplasia, then to malignant cells. A 15-tissue protein panel integrating these signatures demonstrated potential for targeting individuals at high risk for progressing to gastric neoplasia (OR = 7.22, 95% CI: 1.31–39.72 for the high-score group). A 4-circulating protein panel may be used as non-invasive markers predicting the risk of GC development (hazard ratio = 3.73, 95% confidence interval: 1.63–8.54, high-risk *vs.* low-risk populations, area under the curve = 0.75).

Conclusions: Concordant proteomics signatures associated with *H. pylori* infection and gastric carcinogenesis were unveiled with potential as biomarkers for targeted prevention strategies.

KEYWORDS Stomach neoplasms; *Helicobacter pylori*; proteomics; scRNA-seq; biomarker

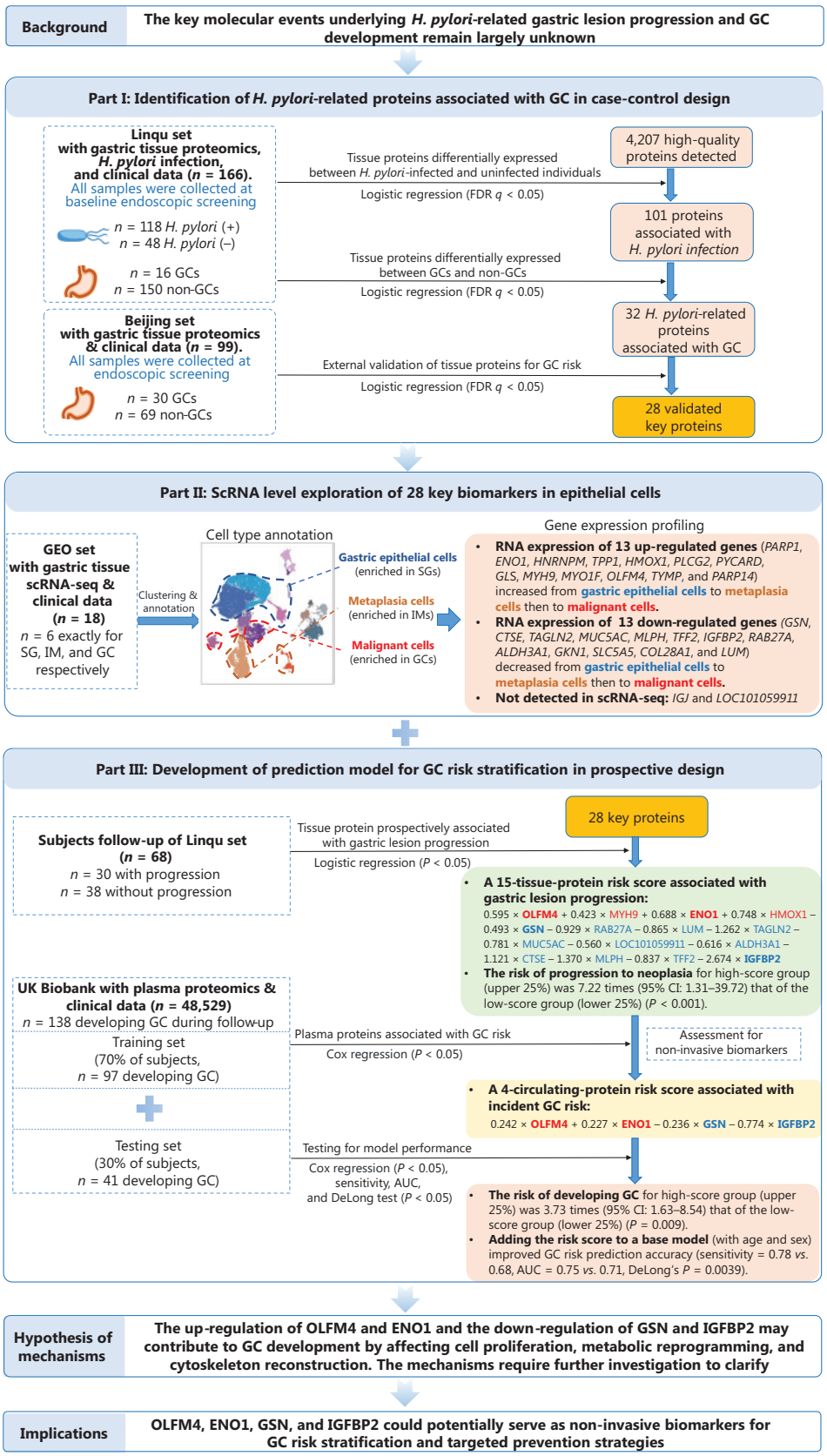
Study flow chart

Part I: The Linqu set of 166 gastric tissues, including 16 GC [11 *H. pylori* (+) and 5 *H. pylori* (−)] plus 150 non-GC [107 *H. pylori* (+) and 43 *H. pylori* (−)], was used in the discovery phase. Among 4,207 proteins obtained by proteomics analysis, 101 were identified as *H. pylori*-associated and this number was narrowed down to 32 when combined with GC and narrowed further to 28 by verification with the Beijing set, including 30 GC *versus* 69 non-GC (no *H. pylori* status available).

Part II: The GSE249874 scRNA-seq dataset [n = 18 (6 each for gastritis, intestinal metaplasia, and GC), *H. pylori* unknown] was downloaded and cell types were annotated for comparison to explore the potential expression profile of the 28 genes during the epithelial transition. Among the 28 genes, 13 were found up- and another 13 down-regulated during the progression from gastric epithelium to intestinal metaplasia and further to malignant cells (2 genes not detected).

Part III: The potential association of the 28 proteins with gastric lesion progression was tested through prospective analysis of the endoscopic follow-up data of the Linqu set (30 and

*These authors contributed equally to this work.
Correspondence to: Wenqing Li
E-mail: wenqing_li@bjmu.edu.cn
ORCID ID: https://orcid.org/0000-0002-8336-5552
Received February 14, 2025; accepted May 26, 2025.
Available at www.cancerbiomed.org
©2025 The Authors. Creative Commons Attribution-NonCommercial 4.0 International License


Background
The key molecular events underlying H. pylori-related gastric lesion progression and GC development remain largely unknown
Part I: Identification of H. pylori-related proteins associated with GC in case-control design
Linqu set with gastric tissue proteomics, H. pylori infection, and clinical data (n = 166).
All samples were collected at baseline endoscopic screening
n = 118 H. pylori (+)
n = 48 H. pylori (–)
n = 16 GCs
n = 150 non-GCs
Beijing set with gastric tissue proteomics & clinical data (n = 99).
All samples were collected at endoscopic screening
n = 30 GCs
n = 69 non-GCs
Tissue proteins differentially expressed between H. pylori-infected and uninfected individuals
Logistic regression (FDR q < 0.05)
4,207 high-quality proteins detected
101 proteins associated with H. pylori infection
Tissue proteins differentially expressed between GCs and non-GCs
Logistic regression (FDR q < 0.05)
32 H. pylori-related proteins associated with GC
External validation of tissue proteins for GC risk
Logistic regression (FDR q < 0.05)
28 validated key proteins
Part II: ScRNA level exploration of 28 key biomarkers in epithelial cells
GEO set with gastric tissue scRNA-seq & clinical data (n = 18)
n = 6 exactly for SG, IM, and GC respectively
Clustering & annotation
Cell type annotation
Gastric epithelial cells (enriched in SGs)
Metaplasia cells (enriched in IMs)
Malignant cells (enriched in GCs)
Gene expression profiling
RNA expression of 13 up-regulated genes (PARP1, ENO1, HNRNPM, TPP1, HMOX1, PLCG2, PYCARD, GLS, MYH9, MYO1F, OLFM4, TYMP, and PARP14) increased from gastric epithelial cells to metaplasia cells then to malignant cells.
RNA expression of 13 down-regulated genes (GSN, CTSE, TAGLN2, MUC5AC, MLPH, TFF2, IGFBP2, RAB27A, ALDH3A1, GKN1, SLC5A5, COL28A1, and LUM) decreased from gastric epithelial cells to metaplasia cells then to malignant cells.
Not detected in scRNA-seq: IGJ and LOC101059911
Part III: Development of prediction model for GC risk stratification in prospective design
Subjects follow-up of Linqu set (n = 68)
n = 30 with progression
n = 38 without progression
Tissue protein prospectively associated with gastric lesion progression
Logistic regression (P < 0.05)
28 key proteins
A 15-tissue-protein risk score associated with gastric lesion progression:
0.595 × OLFM4 + 0.423 × MYH9 + 0.688 × ENO1 + 0.748 × HMOX1 – 0.493 × GSN – 0.929 × RAB27A – 0.865 × LUM – 1.262 × TAGLN2 – 0.781 × MUC5AC – 0.560 × LOC101059911 – 0.616 × ALDH3A1 – 1.121 × CTSE – 1.370 × MLPH – 0.837 × TFF2 – 2.674 × IGFBP2
The risk of progression to neoplasia for high-score group (upper 25%) was 7.22 times (95% CI: 1.31–39.72) that of the low-score group (lower 25%) (P < 0.001).
UK Biobank with plasma proteomics & clinical data (n = 48,529)
n = 138 developing GC during follow-up
Training set (70% of subjects, n = 97 developing GC)
Testing set (30% of subjects, n = 41 developing GC)
Plasma proteins associated with GC risk
Cox regression (P < 0.05)
Assessment for non-invasive biomarkers
A 4-circulating-protein risk score associated with incident GC risk:
0.242 × OLFM4 + 0.227 × ENO1 – 0.236 × GSN – 0.774 × IGFBP2
Testing for model performance
Cox regression (P < 0.05), sensitivity, AUC, and DeLong test (P < 0.05)
The risk of developing GC for high-score group (upper 25%) was 3.73 times (95% CI: 1.63–8.54) that of the low-score group (lower 25%) (P = 0.009).
Adding the risk score to a base model (with age and sex) improved GC risk prediction accuracy (sensitivity = 0.78 vs. 0.68, AUC = 0.75 vs. 0.71, DeLong's P = 0.0039).
Hypothesis of mechanisms
The up-regulation of OLFM4 and ENO1 and the down-regulation of GSN and IGFBP2 may contribute to GC development by affecting cell proliferation, metabolic reprogramming, and cytoskeleton reconstruction. The mechanisms require further investigation to clarify
Implications
OLFM4, ENO1, GSN, and IGFBP2 could potentially serve as non-invasive biomarkers for GC risk stratification and targeted prevention strategies

38 with and without progression, respectively) to develop GC risk models. Logistic regression showed that 15 of the 28 proteins were statistically significant and therefore used to construct a 15-tissue protein risk model using a weighted summing approach (sum of the standardized expression of 15 proteins × the logistic regression coefficient). The risk scores produced by the model were associated with the risk of progression to neoplasia with the high-score group (subjects with scores in the upper quartile) 7.22 times (95% CI: 1.31–39.72) more likely to develop GC compared to the low-score group (subjects with scores in the lower quartile). The UK Biobank dataset (n = 48,529,138 GC) was subsequently used to develop a non-invasive prediction model with the 4 plasma proteins [shown in bold (70% data for training)]. A 4-plasma protein non-invasive GC risk prediction model was established using the same weighting method and tested on the remaining 30% UK data. The risk of developing GC for the high-score group (subjects with scores in the upper quartile) was 3.73 times (95% CI: 1.63–8.54) higher compared to the low-score group (subjects with scores in the lower quartile; P = 0.009). Finally, combining this risk score to the previous model (based on age and gender only) increased the prediction accuracy, with a sensitivity from 0.68–0.78 and an AUC from 0.71–0.75 (DeLong's P = 0.0039).

AUC, area under the receiver operating characteristic curve; CI, confidence interval; FDR, false discovery rate; GC, gastric cancer; *H. pylori*, *Helicobacter pylori*; IM, intestinal metaplasia; scRNA-seq, single-cell RNA sequencing; SG, superficial gastritis

Introduction

Gastric cancer (GC) accounts for a significant percentage of cancer-related deaths worldwide[1]. The progression of intestinal-type GC, a prevalent subtype in China, typically follows a multi-stage sequence known as Correa's cascade[2], which is usually initiated with superficial gastritis (SG) and may evolve into chronic atrophic gastritis (CAG) and develop intestinal metaplasia (IM) thereafter. Furthermore, some cases could progress to low- (LGIN) or high-grade intraepithelial neoplasia (HGIN) and ultimately deteriorate into invasive GC. As a leading cause of gastric carcinogenesis *Helicobacter pylori (H. pylori)* could explain up to 90% of non-cardia GC[3], while eradication of *H. pylori* infection significantly decreases the risk of developing GC according to previous randomized trials[4-6]. Considering the pivotal role of *H. pylori* infection, a thorough investigation into the fundamental molecular events underpinning the *H. pylori* carcinogenic process will provide novel insights into GC etiology and facilitate the risk stratification and early detection of GC.

Despite expanding knowledge on the mechanisms underlying chronic inflammation triggered by *H. pylori*[7,8] and the concerted efforts to explore the GC etiology through metabolic[9], lipid[10], immune[11], and microbial[12] perspectives, a major knowledge gap persists in our holistic understanding of the complex interactions between *H. pylori* infection and the multifaceted processes leading to GC development. Few studies have simultaneously compared the key molecular alterations induced by *H. pylori* infection with the molecular alterations that occur during the progression of gastric lesions. This lack of studies hinders our capacity to attain a deep understanding of the intrinsic link between *H. pylori* infection and GC, and to decipher the core molecular signatures underlying the carcinogenesis process.

To systematically investigate proteomic alterations associated with *H. pylori*-driven gastric carcinogenesis, an integrated multi-omics strategy was adopted by synergizing proteomic profiling with single-cell RNA (scRNA) sequencing by considering the multistage cascade progression of gastric lesions to GC. We then sought to establish robust tissue-protein and circulating-protein risk score models for GC risk stratification based on a prospective follow-up of our in-house cohort and a large publicly accessible dataset. The objective was to gain novel insights into the key molecular signatures that underlie *H. pylori*-related gastric carcinogenesis and to develop a practical protein-based translational strategy for the targeted prevention of GC.

Materials and methods

Study design and subjects

This study utilized a mixed study design integrating case-control studies of proteomic signatures for *H. pylori* infection and risk of gastric lesions, validation of biomarkers at the single-cell level, and prospective studies of key proteins associated with the risk of gastric lesion progression and GC development (**Study flow chart**, **Table S1**).

The differentially expressed proteins between *H. pylori* positive and negative individuals were examined during the case-control study phase by leveraging in-house resources with proteomics profiling of gastric tissues (Linqu set, n = 166 subjects, 118 *H. pylori*-positive *vs.* 48 *H. pylori*-negative; **Table S2**). Details of the Linqu proteomics dataset have been

described previously[13]. The subjects were enrolled from the National Upper Gastrointestinal Cancer Early Detection (UGCED) Program in Linqu county, Shandong province of China, a known high-risk GC area, between 22 November and 7 December 2018. The subjects underwent a questionnaire interview, blood sampling, and endoscopic examination. Sixteen individuals diagnosed with HGIN or invasive GC were classified as the GC group because they shared similar treatment principles within the program. Among the non-GC controls, 80 had advanced pre-cancerous gastric lesions (specifically IM or LGIN) and 70 had mild gastric lesions (specifically SG or CAG). Enzyme-linked immunosorbent assays (ELISAs) were used to determine *H. pylori* infection status through IgG serology. Ninety-four subjects also underwent the ^{13}C-urea breath test (CUBT). In addition, 68 subjects with gastric lesions had repeat endoscopic examinations and histopathologic diagnoses during the follow-up period until 10 October 2023 with information on the progression of gastric lesions documented. Whether *H. pylori*-related proteomic signatures were also associated with GC compared to individuals without GC in the Linqu set (n = 166) and the independent Beijing set (n = 99) were determined. The Beijing set included 30 patients with GC enrolled from Peking University Cancer Hospital and 69 patients with non-GC gastric lesions (24 with mild gastric lesions and 45 with advanced gastric lesions) enrolled from Dongfang Hospital in Beijing (**Table S3**). The in-house proteomic datasets are deposited in the ProteomeXchange Consortium *via* the iProX partner repository (access No. IPX0003438002, DOI: 10.1016/j.ebiom.2021.103714). Gene expression patterns at the single-cell level for key proteins validated for GC risk were further investigated based on a publicly available scRNA dataset from the GEO database (n = 18, access No. GSE249874, DOI: 10.1136/gutjnl-2023-iddf.110), which is comprised of 6 subjects with gastritis, 6 with intestinal metaplasia, and 6 with GC[14].

The potential association between the baseline expression of key tissue proteins and the risk of gastric lesion progression was investigated during the prospective study phase utilizing prospective endoscopic follow-up of 68 participants in the Linqu set. A tissue protein risk score was calculated based on progression-related proteins. Additionally, the relationship between the circulating levels and the risk of incident GC was explored, leveraging the UK Biobank (UKB, n = 48,529). The efficacy of a panel of key circulating proteins in stratifying risk populations and predicting the risk of developing GC was also evaluated. A circulating protein risk score was established and the prediction potential for GC risk was examined in the UKB. The UKB is a large population-based cohort of UK residents 40–59 years of age who were recruited between 2006 and 2010[15]. For this study, 48,529 subjects with proteomic data and without a cancer diagnosis at cohort enrollment were selected, of whom 138 developed GC during the follow-up period. Information on *H. pylori* infection status was not available for the Beijing set, scRNA dataset, or the UKB.

The study was approved by the Institutional Review Board of Peking University Cancer Hospital (No. 2018KT117). All participants of the Linqu set and Beijing set provided written informed consent.

Gastroscopy and histopathology

Gastroscopic examinations of the Linqu and Beijing sets were performed by senior gastroenterologists utilizing video endoscopes (GIF-260; Olympus, Tokyo, Japan). The biopsies were fixed with formalin, embedded in paraffin, then evaluated independently by two pathologists adhering to the guidelines set by the Chinese Association of Gastric Cancer and the Updated Sydney System. Each specimen was diagnosed as normal, SG, CAG, IM, LGIN, HGIN, or invasive GC. A global diagnosis was subsequently assigned to each subject based on the most severe diagnosis identified from all the biopsies. To prevent systemic biases in proteomic profiles stemming from discrepancies between global and site-specific diagnoses, participants exhibiting the most severe histologic manifestations in the gastric mucosa at the lesser curvature of the antrum or angulus were exclusively enrolled, from which biopsies were collected for both histopathologic and proteomic assays. Each subject was given a histopathology severity score at baseline and at the conclusion of the follow-up period. Progression of gastric lesions was deemed to have occurred if the endpoint score exceeded the baseline score for a given subject.

Proteomics quantification

The tissue proteomic assay was determined by liquid chromatography-tandem mass spectrometry, as described previously[13]. Standard protocols were followed for data processing and protein quantitation. Specifically, the Firmiana pipeline was used to analyze raw mass spectrometry data[16]. Protein identification was carried out using the Mascot search engine (Matrix Science, version 2.3) against the NCBI human Refseq protein database (dated 04/07/2013). Protein abundance was

quantified using a label-free intensity-based absolute quantification (iBAQ) method and these values were converted to intensity-based fractions of the total (iFOT). The log10-transformed and normalized iFOT values for each protein were used for further analysis. Quality control assessments confirmed the stability of the proteomic profiles and all samples exhibited consistent quantitation of the tissue proteome. A total of 4,207 proteins that were identified in at least 25% of the samples were included.

Olink Explore 3,072 proteomic profiling was performed on EDTA-plasma samples from approximately 54,000 participants in the UKB[17]. Olink utilizes the proximity extension assay. Microfluidic quantitative PCR was subsequently used for relative quantification and analysis of the amplified DNA tags. Details of the detection method and quality control procedures have been described elsewhere[15]. Protein levels undergo intra- and cross-batch normalization processes and are ultimately expressed as normalized protein expression.

Single-cell RNA sequencing

Preprocessing of raw scRNA-seq data followed the CellRanger pipeline (version 7.2.0) using the GRCh38 as the reference genome[18]. Downstream analyses of scRNA-seq data followed the Seurat pipeline (version 5.1.0)[19]. Specifically, cells with > 50,000 unique molecular identifier counts, <300 or >7,500 detected genes, >5% hemoglobin gene counts, or >30% mitochondrial gene counts were filtered out. Potential doublet cells were recognized and filtered out following the DoubletFinder pipeline (version 2.0.4).

The *NormalizeData*, *FindVariableFeatures*, and *ScaleData* functions in Seurat were utilized to standardize library sizes, perform logarithm transformation, and identify and scale 2,000 highly variable genes (HVGs). Harmony (version 1.2.3) was used to correct for batch effects with default parameters prior to dimensionality reduction. *RunPCA*, *FindNeighbors*, *FindClusters*, and *RunUMAP* functions in Seurat were then used to extract the top 30 principal components to construct a shared nearest neighbor graph, perform Louvain unsupervised clustering, and visually represent the results utilizing the scaled HVG expression matrix.

Bulk RNA-seq quantification

The normalized bulk RNA-seq data generated by The Cancer Genome Atlas (TCGA) on primary stomach adenocarcinoma (STAD) from the NCI Cancer Genomic Data Commons (NCI-GDC: https://gdc.cancer.gov) was downloaded. The RNA-seq data was processed and normalized by the NCI-GDC bioinformatics team using the transcriptome analysis pipeline.

Bioinformatics and statistical analyses

Analyses were performed using R 4.4.1 unless otherwise stated.

Proteomic signature for H. pylori infection and GC

The proteomic data underwent initial processing through principal component analyses (PCA) utilizing top two principal components for visualization of *H. pylori*-positive and -negative individuals, as well as individuals with GC, advanced gastric lesions, and mild gastric lesions. The *stat_ellipse* function from the ggplot2 package (version 3.5.1) was subsequently used to depict grouped ellipses at a 95% confidence level. A heatmap with hierarchical clustering was generated using the pheatmap package (version 1.0.12). Gene ontology (GO) enrichment analysis was used to identify *H. pylori*-related pathways using the clusterProfiler (version 4.12.6) and org.Hs.eg.db packages (version 3.18.0).

Unconditional logistic regression model, as is often recommended for case-control studies[20], was used to identify differentially expressed proteins between *H. pylori*-positive and -negative individuals based on the Linqu set, incorporating adjustments for age, gender, and gastric histopathology. Proteins with a two-sided false discovery rate (FDR)-q <0.05 were considered statistically significant. Upregulated proteins in *H. pylori*-positive individuals had an odds ratio (OR) >1, whereas downregulated proteins had an OR <1. Primary analyses were conducted using *H. pylori* serology. Additionally, association analyses were performed for significant proteins and correlated with the infection status as determined by CUBT alone or in combination with serology.

Significant *H. pylori* infection-related proteins were further examined for associations with GC and compared to non-GC based on the Linqu and Beijing sets. ORs and corresponding confidence intervals (CIs) were calculated by unconditional logistic regression models adjusting for *H. pylori* infection (for Linqu set only), age, and gender. Those with an FDR-q <0.05 in both sets were highlighted as key proteins associated with *H. pylori* infection and GC and were visualized by forest plot. The Kruskal-Wallis rank-sum test was subsequently used as a complementary approach to pinpoint proteins exhibiting

expression changes throughout the progression of gastric lesions, spanning from mild-to-advanced gastric lesions and ultimately to GC.

Spearman correlation analyses were performed to compare the association ORs of these proteins for *H. pylori* with GC to investigate whether individual proteins exhibit consistent associations with *H. pylori* and GC. The comparison of *H. pylori*- and GC-related pathways was facilitated by a two-dimensional GO annotation utilizing significant proteins associated with *H. pylori* infection and GC[21]. Similarly, Spearman correlation analyses were also performed to compare the enrichment scores of GO terms for *H. pylori* to GC.

Ensemble learning (e.g., stacking models), while offering potential advantages in integrating diverse models and enhancing overall model performance[22], may carry limitations, such as potential bias in effect estimation[23] and reduced biological interpretability, compared to traditional statistical models[24]. Moreover, ensemble learning strategies are prone to overfitting when applied to a modest sample size[25]. Given the objectives and typical characteristics of molecular epidemiologic research, ensemble learning approaches were not adopted for this and below analyses.

Quantification of key protein gene expression at the scRNA level

The FindAllMarkers function in the Seurat R package (version 5.1.0) was used to identify the top 30 marker genes for each cluster and ranked by fold change based on the results of unsupervised clustering and the uniform manifold approximation and projection. These markers were then cross-referenced with published literature and cell-type-specific databases to annotate the major cell lineages, which included epithelial cells (encompassing endocrine cells), stromal cells (normal fibroblasts, cancer-associated fibroblasts, endothelial cells, and pericytes), immune cells (myeloid cells, mast cells, T cells, NK cells, B cells, and plasma cells), and proliferating cells. Bubble plots were created for the curated list of canonical cell markers following the methodology described in previous reports.[11,26]. The patterns of expression across major cell types were examined utilizing key protein biomarkers. The proportion of corresponding RNA expressed across cell types in the scRNA dataset for each candidate protein were calculated and plotted in a proportional heatmap.

Considering the epithelial-origin nature of GC, epithelial and endocrine cells were further selected *in silico* for the following scRNA-analyses in a two-tiered clustering strategy. Epithelial and endocrine clusters were manually annotated and retained based on canonical markers. A more refined clustering process was subsequently carried out leveraging cluster-specific genes identified through the *FindAllMarkers* function in Seurat, as well as markers documented in the literature and databases[27-31]. Clusters that exhibited concurrent high expression of lineage makers unrelated to epithelial and endocrine cells during the second round were considered putative doublets and excluded from subsequent analyses. The *DotPlot* function in Seurat was used to quantify the levels and proportions of specified genes in normal, IM, and malignant epithelial cells.

The signature scores of selected genes were calculated for *H. pylori* and GC up- and down-regulated proteins using the *AddModelScore* function in Seurat. The signature scores across epithelial and endocrine cell types were visualized by *VlnPlot* function in Seurat.

Identification of tissue proteins associated with progression of gastric lesions

Whether baseline expression levels of specific tissue proteins correlated with the progression of gastric lesions over time were investigated by conducting endoscopic follow-up evaluations on 68 individuals from the Linqu cohort. Unconditional logistic regression models adjusted for factors, such as age, gender, *H. pylori* infection status, and baseline histopathology, were used to compute ORs (CIs) for these associations. Logistic regression analysis was appropriate due to the short, uniform follow-up period (mean = 1.5 years) based on scheduled endoscopies, which caused imprecise timing and interval censoring bias that limits time-to-event analysis. Statistical significance was established at a single-sided P-value <0.05 for the prospective association analyses. A tissue protein risk score was formulated for tissue proteins that consistently demonstrated a risk for GC in the Linqu and Beijing cohorts, as well as the progression risk of gastric lesions in the Linqu follow-up. This score was derived by summing the weighted expressions of each protein, in which the weight was determined by multiplying the standardized baseline expression of protein by the logistic regression coefficient obtained. The subjects were stratified into the following three risk categories: low (from 0–25% of the tissue protein risk score); medium (from 25%–75%); and high (from 75%–100%). The association of this tissue protein risk score was then investigated with an alternative outcome of "progression to gastric neoplasia," which was exclusively defined as progression to gastric neoplasia, encompassing LGIN or more severe conditions, for individuals initially with IM or less severe gastric lesions at baseline.

The analysis was further extended to the bulk transcriptomic level to validate the tissue protein risk score. A tissue RNA risk score was computed based on the TCGA-STAD dataset by applying the protein score weighting coefficients to bulk RNA-seq data and the distribution between tumor and adjacent normal tissues was compared.

Identification of circulating proteins associated with risk of incident GC

Given the considerable value of non-invasive biomarker testing, whether the plasma levels of these proteins exhibited an association with the risk of incident GC during the prospective follow-up was examined by leveraging data from the UKB. Cox proportional hazards regression models were used to calculate age- and gender-adjusted hazard ratios (HRs) and CIs for associations utilizing 70% of the UKB samples, which is aligned with methodologic recommendations[20]. The proportional hazards assumption was tested through Schoenfeld residual testing and no violations were detected. A circulating protein risk score model was developed for proteins that consistently showed a relationship with progression of gastric lesions during the Linqu set follow-up and the occurrence of GC in the UKB by leveraging the precise event timing to provide an accurate estimation of the time-to-GC risk. This model was created by summing the weighted expression of each protein, where the weight was determined by multiplying the baseline standardized expression of the protein by the regression coefficient derived from Cox proportional hazards regression. Subjects were also stratified into 3 risk groups based on this circulating protein risk score, as follows: low (from 0%–25% of the circulating protein risk score); medium (from 25%–75%); and high (from 75%–100%).

Construction of the circulating protein-based risk prediction model

Logistic regression models were then used to develop risk prediction models for GC, incorporating both age and gender independently and in conjunction with the risk score of circulating proteins. Considering the category imbalance between GC and non-GC cases in the UKB, a combination-sampling approach with 10-fold cross-validation was used [ROSE (version 0.0.4) and caret (version 6.0.94) packages][32]. Non-GC controls were randomly down-sampled to 138 to establish a 1:1 ratio with GC patients. Models were trained on 70% of the UKB subjects and the performance was validated in the remaining 30%. The sensitivity, specificity, and utilized area under the receiver operating characteristic curve (AUC) were calculated and the DeLong test was used to compare the performance of prediction models.

Results

Characteristics of the study participants

The characteristics of participants from the Linqu set are shown in **Table S2**. Of the 166 subjects, 118 were *H. pylori*-seropositive based on serologic testing. Among the 94 subjects tested by CUBT, 40 were *H. pylori*-positive. No significant differences were noted in age distribution, gender, or baseline histopathologic findings between the *H. pylori*-seropositive and -seronegative groups. During the subsequent follow-up, 30 of 68 subjects had gastric lesion progression. The characteristics of participants from the Beijing set, scRNA set, and the UKB are shown in **Table S3**. The mean ages of the Linqu, Beijing, and scRNA sets and the UKB are 56.6, 60.5, 54.2, and 56.5 years, respectively. The proportion of males was slightly higher than females in all datasets but this difference was not statistically significant.

Proteomic profiles according to *H. pylori* infection and multi-stage gastric lesions

Tissue proteomic profiling of the Linqu cohort (n = 166) revealed an overall distinction between proteomic features in individuals with GC and mild (SG/CAG) or advanced gastric lesions (IM/LGIN; **Figure 1A**, top). However, the visual separation between *H. pylori*-positive and -negative groups was not definitive based on the PCA plot (**Figure 1A**, bottom). Hierarchical clustering analysis revealed that proteins enriched in *H. pylori*-seropositive samples were characterized by an aggregation among individuals with GC or advanced gastric lesions [IM/LGIN (highlighted in the lower part, **Figure 1B**)]. In contrast, the protein pattern in samples from *H. pylori*-seronegative patients exhibited a similar aggregation pattern to individuals with mild gastric lesions [SG/CAG (highlighted in the upper part, **Figure 1B**)]. Subjects who had consistent test results on *H. pylori* infection by serologic testing and CUBT tended to cluster together. The proteomic feature differences among various histopathologic groups were highly concordant with the status variations among different *H. pylori* infections, suggesting that *H. pylori* infection may be a significant factor influencing the protein signatures linked to the sequential development of GC.

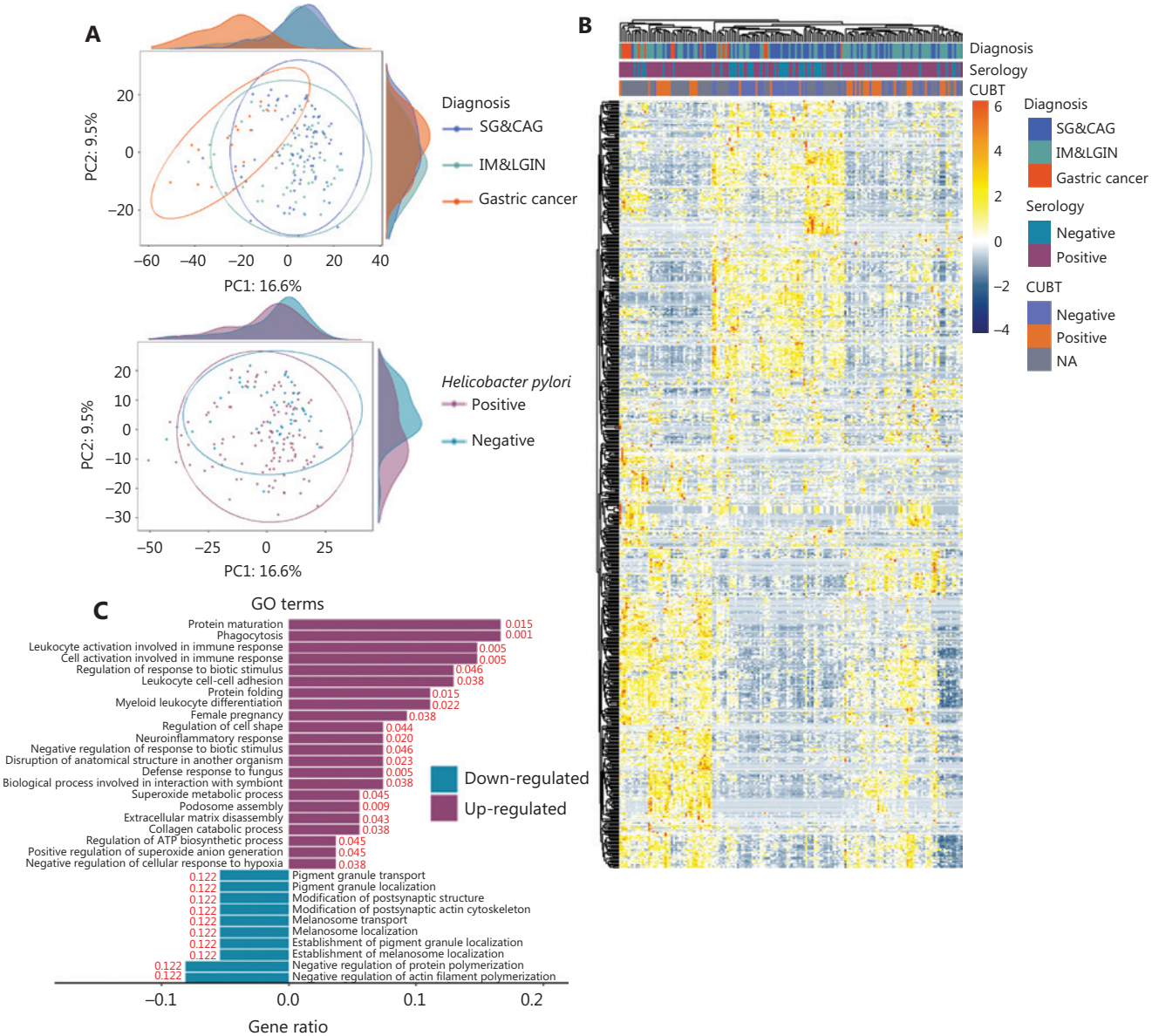


Figure 1 Proteomics profiling according to *H. pylori* infection and multi-stage gastric lesions. A. Principal component analyses for dimensional reduction of proteomic profiles according to histopathologic diagnosis (top) and *H. pylori* infection status (bottom); B. Hierarchical clustering for proteomics pattern according to gastric histopathology and *H. pylori* infection as determined by serology or CUBT; C. Gene ontology enrichment analyses of *H. pylori*-related proteins. FDR of each term is labeled in red; PC, principal component; CUBT, ^{13}C-urea breath test; *H. pylori*, *Helicobacter pylori*; SG, superficial gastritis; CAG, chronic atrophic gastritis; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; FDR, false discovery rate.

H. pylori-seropositive and -seronegative individuals within the Linqu cohort were compared and 101 proteins were differentially expressed with an FDR-q <0.05. The infection status, as ascertained by CUBT alone and in combination with serologic testing, was also utilized to test the robustness of findings. Highly consistent lists of differentially expressed proteins were identified based on the three methods used to determine *H. pylori* infection status (**Table S4**). GO enrichment analyses of 101 proteins revealed that the *H. pylori*-upregulated proteins are primarily involved in biological processes, such as protein folding and maturation, phagocytosis, leukocyte activation, differentiation, adhesion, and the immune response, suggesting that *H. pylori* infection is associated with activation of immune responses and synthesis of some proteins (**Figure 1C**).

Proteins significantly associated with *H. pylori* infection and GC risk

Whether 101 *H. pylori*-related proteins are associated with the risk of GC was determined to offer insights into the potential influence of *H. pylori* infection on the development of GC. This analysis yielded 32 statistically significant proteins (both FDR-q <0.05 for each protein), 28 of which were successfully validated for GC risk in an independent dataset, the Beijing cohort (n = 99), including 15 down- and 13 up-regulated proteins (**Figure 2A**). Proteins elevated in *H. pylori*-positive subjects, such as OLFM4, PYCARD, TYMP, and ENO1, were also increased in patients with GC, while decreased proteins, like IGFBP2 and TFF2, had lower expression. Spearman correlation analysis revealed a significant correlation between *H. pylori*- and GC-associated protein alterations ($\rho = 0.455$, $P = 6.68 \times 10^{-7}$). Despite this significant correlation, many proteins exhibited an asymmetric dispersion pattern relative to the correlation curve, particularly those proteins that were downregulated. In addition, these proteins also exhibited a trend in progression from mild-to-advanced gastric lesions to GC, which is consistent with the direction of the associations with *H. pylori* infection and GC risk (**Figures S1 and S2**). Therefore, these proteins were defined as "key markers" associated with *H. pylori*-related gastric lesion progression and GC development.

A strong positive correlation was demonstrated between the enrichment scores of GO terms related to *H. pylori* infection and GC by applying GO enrichment analysis to these proteins ($\rho = 0.784$, $P = 1.80 \times 10^{-36}$; **Figure 2C**). Notably, the pathways involving protein production, regulation of translation, amide metabolic processes, and glycolytic processes were upregulated by *H. pylori* infection and in GC, while downregulated terms were primarily related to hormone and ethanol metabolic processes.

ScRNA localization of key markers revealed gastric epithelial carcinogenesis

Building on these 28 key proteins, the cellular subtypes in which potential regulatory mechanisms driving *H. pylori*-induced gastric carcinogenesis might be operative were identified. To this end, scRNA-seq analysis was performed on a publicly available dataset (n = 18), which encompasses histopathology samples collected from multiple stages of gastric lesions. The whole single-cell mapping contained 135,082 high-quality cells spanning 11 major cell types (**Figure 3A, B**). The genes corresponding to the *H. pylori* and GC-downregulated markers were primarily expressed in epithelial cells, whereas upregulated markers were commonly expressed in epithelial cells, myeloid immune cells, lymphocytes, and stromal cells (**Figure 3C**). This finding aligns with our hypothesis that *H. pylori* infection could elicit immune responses and subtly modify the gastric microenvironment, thereby underlying gastric carcinogenesis. Additionally, the enrichment of downregulated markers specifically in the epithelial cells may suggest pivotal changes occurring within this cell type.

Recognizing the epithelial-origin of GC, the expression profiles of the genes for key proteins in epithelial cells were analyzed. The final annotated single-cell mapping contained 29,047 high-quality cells. Unsupervised clustering followed by uniform manifold approximation and projection visualization identified 15 fine-grained cell clusters (**Figure 3D**). The top differentially expressed genes across these cell clusters are shown in **Table S5** and **Figure S3**. A normal gastric epithelial (pit mucous, gland mucous, chief, and parietal cells), IM (enterocyte- and enteroendocrine-like cells), and malignant cell group (malignant-like C1–3) were further defined based on *a priori* biological knowledge and cluster-specific genes. Other cell types included a proliferating cell cluster marked by *MKI67* (proliferative cells), a cluster marked by *POSTN* expression (*POSTN*$^+$ cells), and four endocrine clusters (gastric endocrine, neuroendocrine, *TPH1*$^+$ endocrine, and *NALCN*$^+$ endocrine cells). The marked areas of the normal gastric epithelial, IM, and malignant cell group corresponded well with the histopathologic diagnosis of cell source samples (**Figure 3E**), indicating an epithelial transition during the multi-stage progression of gastric lesions-to-GC.

The transcription levels of "key marker" genes during the epithelial transition were evaluated. The downregulated key marker genes (e.g., *MUC5AC*, *MLPH*, *IGFBP2*, and *ALDH3A1*) were predominantly expressed in normal gastric epithelial cells but remarkably decreased in IM and malignant cells (**Figure 3F**). As an exception, expression of stromal markers (*COL28A1* and *LUM*) was increased in a subgroup of malignant cells (**Figure S4**), potentially indicating an epithelial-mesenchymal transition. In contrast, the expression of upregulated markers, which were transcribed at a low level in normal gastric cells, increased considerably during the epithelial transition. *PARP14*, *TYMP*, *OLFM4*, and *MYO1F* expression was upregulated in a stepwise manner from normal-to-IM cells, then to malignant epithelial cells. Furthermore, the combined signature scores for down- and up-regulated markers

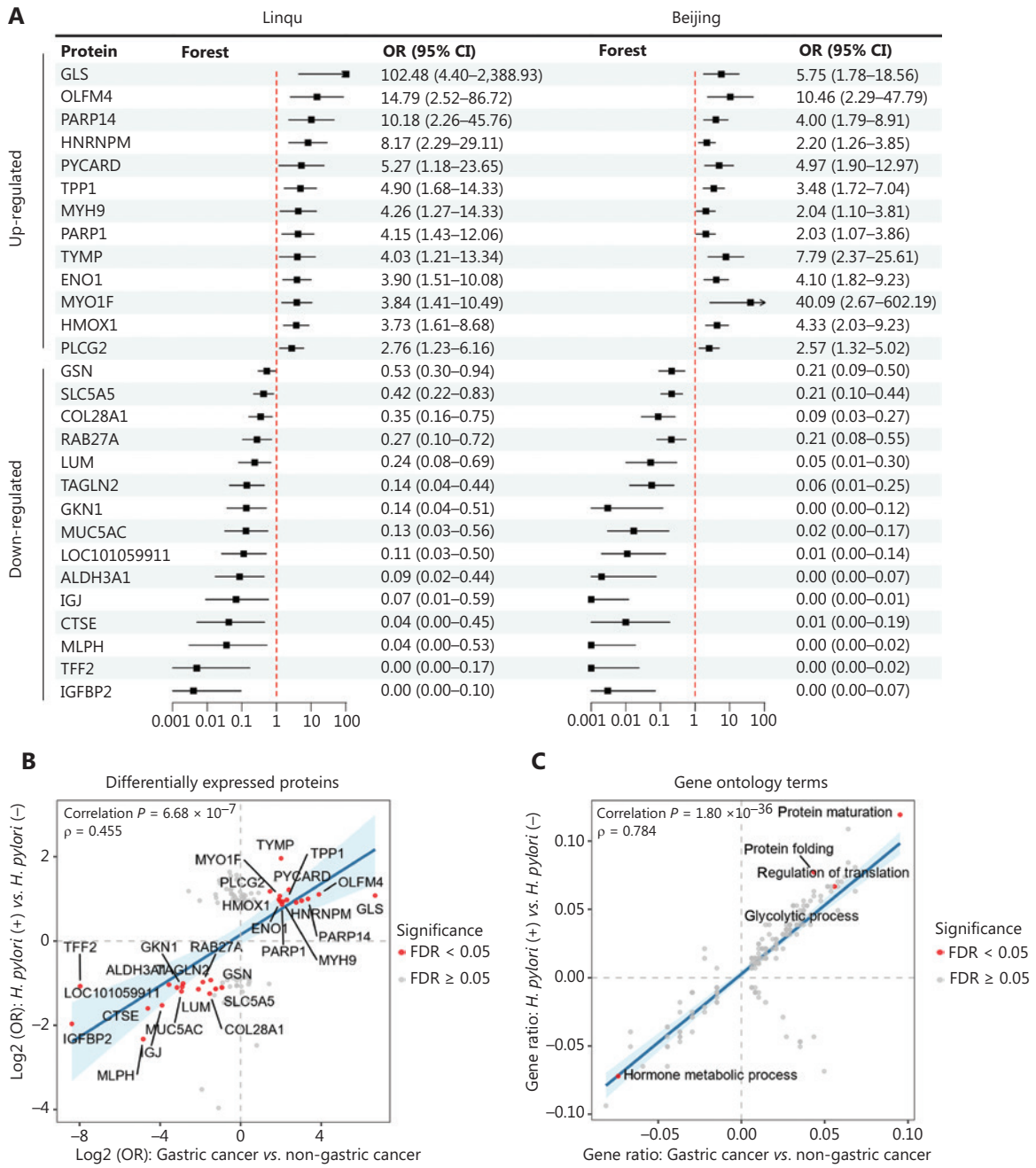


Figure 2 Proteins significantly associated with *H. pylori* infection and gastric cancer risk. A. The associations of 28 *H. pylori*-related proteins (15 down- and 13 up-regulated) with the risk of gastric cancer in the Linqu (left) and Beijing sets (right). ORs (95% CIs) were calculated by unconditional logistic regression adjusting for *H. pylori* serology (in Linqu set), age, and gender. All listed proteins were significantly associated with *H. pylori* serology and gastric cancer at an FDR-*q* <0.05. B. Two-dimensional visualization of the 28 proteins based on ORs associated with gastric cancer (x-axis) and *H. pylori* serology (y-axis). The correlation coefficient and *P* value were calculated using Spearman correlation analysis. C. Two-dimensional visualization of gene ontology terms based on proteins associated with gastric cancer (x-axis) and *H. pylori* serology (y-axis). Terms significantly enriched at an FDR-*q* <0.05 are labeled in red. The correlation coefficient and *P* value were calculated using Spearman correlation analysis. *H. pylori* (−), *Helicobacter pylori*-negative; *H. pylori* (+), *Helicobacter pylori*-positive; OR, odds ratio; CI, confidence interval; FDR, false discovery rate.

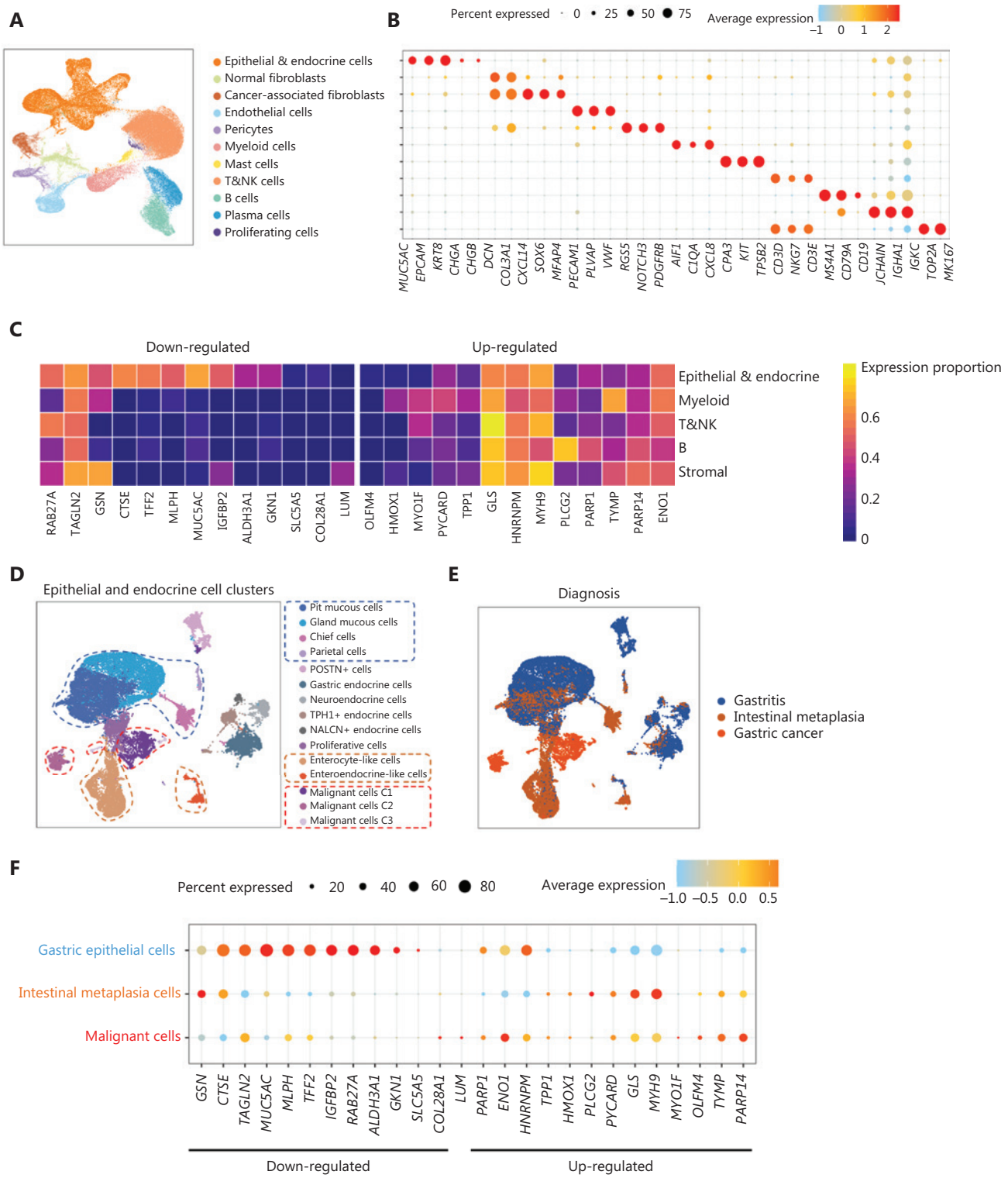


Figure 3 Continued

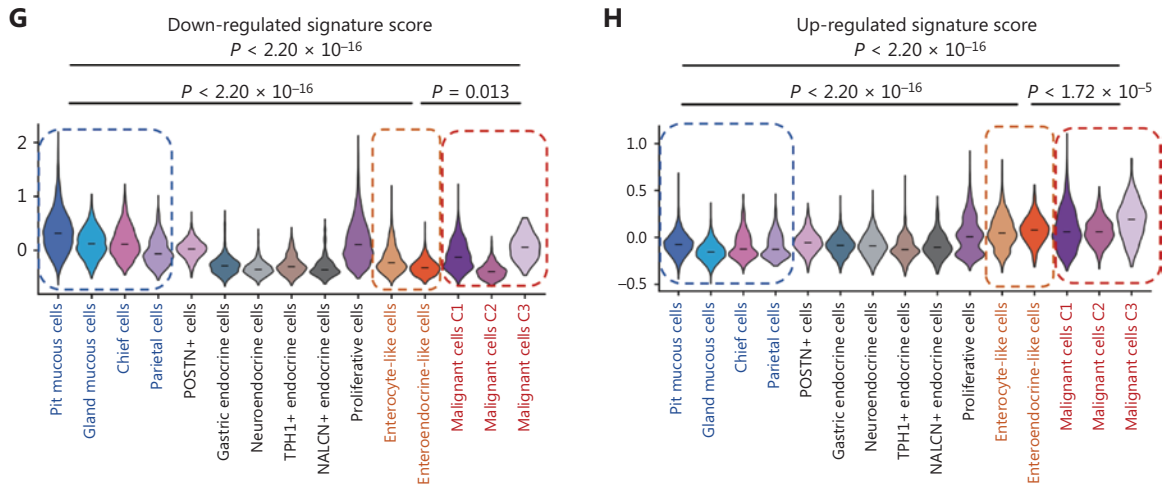


Figure 3 Single-cell RNA characterization of key markers based on the GSE249874 dataset. A. The uniform manifold approximation and projection visualization of total cell populations. B. Bubble plot on the expression of marker gene in each cell cluster. C. Proportional heatmap on the expression of 26 key markers (13 down- and 13 up-regulated) across major cell lineages. Two (IGJ and LOC101059911) of the 28 key proteins significantly associated with *H. pylori* infection and gastric cancer risk lack single-cell RNA expression data and were thus omitted from this and the subsequent single-cell RNA analysis. D. The uniform manifold approximation and projection visualization of 15 epithelial and endocrine cell clusters. Cell types are annotated by canonical markers and cluster-specific markers. Dashed lines are used to encircle the cell clusters of gastric epithelium (colored blue), intestinal metaplastic epithelium (colored brown), and malignant epithelium (colored red). E. The uniform manifold approximation and projection visualization for histopathologic diagnosis based on the tissue origin of the cells. Histopathologic diagnoses are labeled as gastritis (colored blue), intestinal metaplasia (colored brown), or gastric cancer (colored red). F. Bubble plot on the expression of 26 key markers in the cell clusters of gastric (colored blue), intestinal metaplastic (colored brown), and malignant epithelium (colored red). G. The signature score of 13 down-regulated markers across 15 epithelial and endocrine cell clusters. Dashed lines are used to encircle the cell clusters of gastric (colored blue), intestinal metaplastic (colored brown), and malignant epithelium (colored red). H. The signature score of 13 up-regulated markers across 15 epithelial and endocrine cell clusters. Dashed lines are used to encircle the cell clusters of gastric (colored blue), intestinal metaplastic (colored brown), and malignant epithelium (colored red). Signature scores (E and F) were calculated for *H. pylori*- and GC-up- proteins and -down-regulated proteins, respectively, using the *AddModelScore* function in Seurat.

were calculated (**Figure 3G, H**). The decreasing down-and increasing up-regulated signature score trends with the malignant progression of cell types suggested that the "key markers" could distinguish the malignant potential of epithelial cells ($P < 2.20 \times 10^{-16}$ for normal *vs*. malignant cells).

Protein signatures signifying gastric lesion progression and GC risk

Whether the baseline levels could predict the risk of gastric lesion progression and GC occurrence was examined based on 28 key proteins through prospective follow-up of the Linqu cohort and UKB participants. Endoscopic follow-up examinations were performed on 68 subjects in the Linqu cohort with 30 (44.1%) individuals progressing to more severe gastric lesions during a mean follow-up of 1.5 years (SD: 1.1 years). The expression of 15 tissue proteins was significantly linked to the risk of gastric lesion progression (**Figure 4A**, left). Specifically, 4 upregulated proteins (OLFM4, MYH9, ENO1, and HMOX1) were positively associated with the risk of gastric lesion progression. Conversely, 11 proteins (GSN, RAB27A, LUM, TAGLN2, MUC5AC, LOC101059911, ALDH3A1, CTSE, MLPH, TFF2, and IGFBP2) demonstrated decreased expression with increasing severity of gastric lesions and GC and were inversely associated with gastric lesion progression risk. Integrating these 15 proteins into a "tissue panel," a weighted risk score was calculated, as follows:

$$\begin{aligned}\text{Risk score} = {} & 0.595 \times \text{OLFM4} + 0.423 \times \text{MYH9} + 0.688 \\ & \times \text{ENO1} + 0.748 \times \text{HMOX1} - 0.493 \times \text{GSN} \\ & - 0.929 \times \text{RAB27A} - 0.865 \times \text{LUM} - 1.262 \\ & \times \text{TAGLN2} - 0.781 \times \text{MUC5AC} - 0.560 \\ & \times \text{LOC101059911} - 0.616 \times \text{ALDH3A1} \\ & - 1.121 \times \text{CTSE} - 1.370 \times \text{MLPH} - 0.837 \\ & \times \text{TFF2} - 2.674 \times \text{IGFBP2}\end{aligned}$$

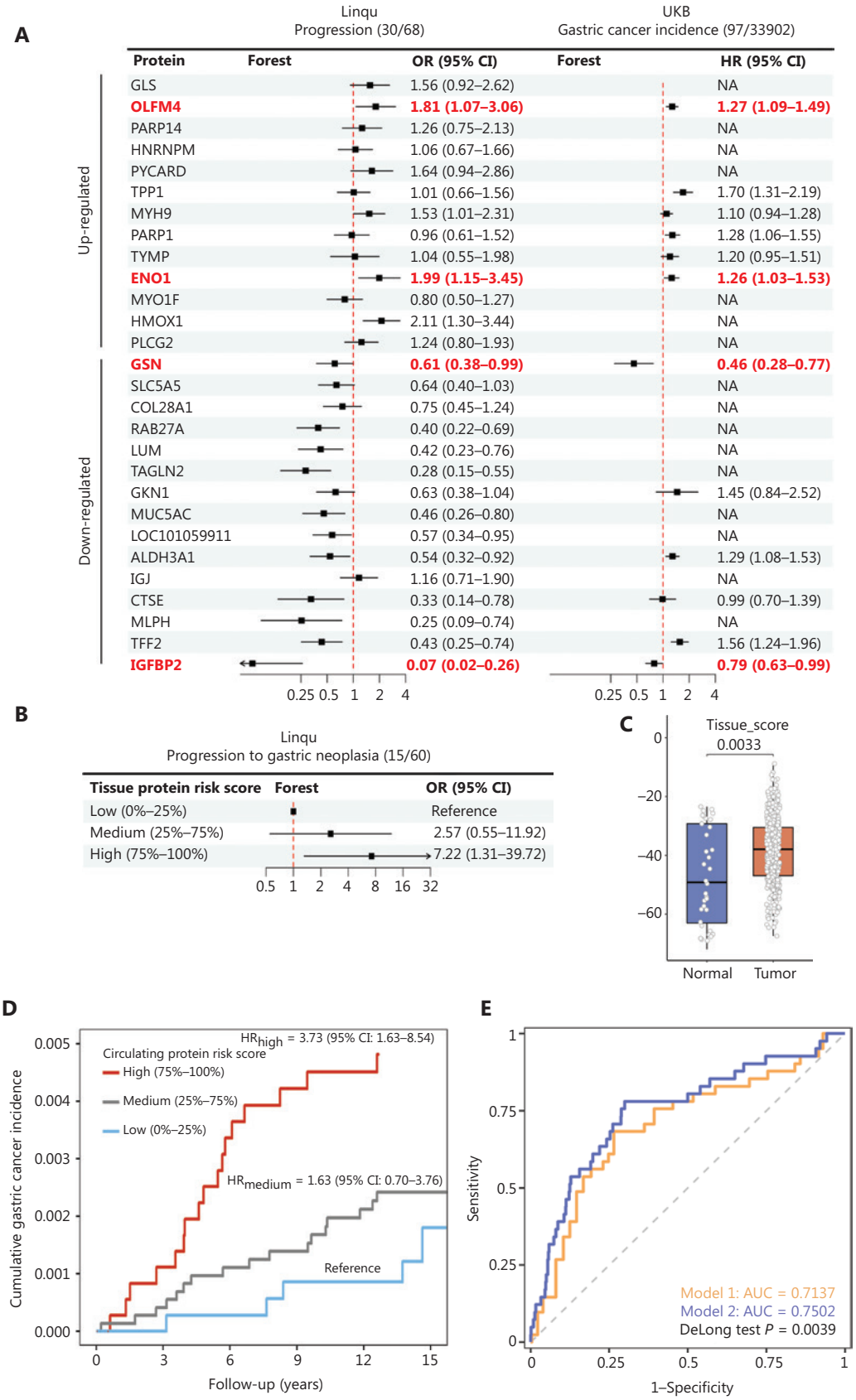


Figure 4 Continued

Figure 4 Protein signatures signifying the risk of gastric lesion progression and GC development. A. Prospective associations of key markers with gastric lesion progression based on follow-up of the Linqu set (left) with incident gastric cancer based on 70% of the UKB as the training set (right). The 15 proteins significantly associated with the risk of gastric lesion progression are bolded. Among the 15 proteins, 4 significantly associated with the risk of GC development are colored in red. B. The association between tissue protein risk scores and the risk of progression to gastric neoplasia. Progression to gastric neoplasia was defined as progression to low-grade intraepithelial neoplasia or more severe conditions for individuals initially with intestinal metaplasia or less severe gastric lesions at baseline. C. RNA level validation of the 15 tissue signature risk score based on the TCGA-STAD dataset. The risk score distribution was compared between GC tumors and adjacent normal samples. D. Cumulative incidence of gastric cancer stratified by circulating protein risk scores calculated using logistic regression based on the remaining 30% of the UKB as the testing set; the low-risk group set was the reference. E. Receiver operating characteristic curves for gastric cancer risk prediction models developed using logistic regression based on the remaining 30% of the UKB as the testing set. Model 1: age and gender; model 2: age, gender, and circulating protein risk score. AUC, area under the curve; CI, confidence interval; HR, hazard ratio; NA, not applicable; OR, odds ratio.

An elevated "tissue protein risk score" was associated with a heightened risk of developing gastric neoplasia (OR = 7.22, 95% CI: 1.31–39.72 for the high score group, **Figure 4B**). The risk score of these 15 tissue signatures was also validated at the RNA level based on the TGCA-STAD dataset. The risk scores for GC tumors were significantly elevated compared to adjacent non-GC tissues (**Figure 4C**, $P = 0.0033$).

Next, the efficacy of applying these proteins as non-invasive markers was assessed for predicting GC risk based on a large-scale prospective cohort of the UKB with plasma proteomic data (n = 48,529), which covered 12 of the 28 selected proteins. During a mean follow-up of 14.5 years (SD: 2.3 years), 138 individuals developed GC. Using 70% of the UKB subjects as the training set, 4 proteins (IGFBP2, GSN, ENO1, and OLFM4) were consistently identified in which circulating levels were associated with the risk of gastric lesion progression and development of GC (**Figure 4A**, right). A weighted risk score was calculated by integrating these 4 proteins as a "circulating panel:"

$$\text{Risk score} = 0.242 \times \text{OLFM4} + 0.227 \times \text{ENO1} - 0.236 \times \text{IGFBP2} - 0.774 \times \text{GSN}$$

The crude incidence rate of GC per 100,000 person-years was 10, 17, and 38 for individuals with a low, medium, and high circulating protein score, respectively (**Figure 4D**), by using the remaining 30% of the UKB as a testing set. The quartile reference ranges of the tissue and circulating protein risk score are shown in the **Table S6**. The hazard ratio (95% CI) of developing GC was 3.73 (1.63–8.54) for the high score group compared to the low circulating protein risk score. Adding the risk score to a base model (only with age and gender) improved the GC risk prediction accuracy (sensitivity = 0.78 *vs*. 0.68, specificity = 0.70 *vs*. 0.69, AUC = 0.75 *vs*. 0.71, Delong's $P = 0.0039$; **Figure 4E**).

Discussion

This is the first study to delve into the key molecular events underlying *H. pylori*-related carcinogenesis by exploring the shared proteomic signatures associated with *H. pylori* infection and the multistage progression of gastric lesions. The molecular signatures potentially driving the malignant transformation of gastric epithelium were uncovered and potential regulatory pathways were pinpointed by incorporating tissue proteomics with scRNA analyses. Tissue and circulating panels that may predict the risk of gastric lesion progression and GC development were established by leveraging key proteins as markers.

H. pylori colonization causes damage to the gastric mucosa, which leads to the development of chronic gastritis, a well-established precursor in the carcinogenic process. Thus, the current study was initiated with *a priori* focus on *H. pylori*-induced proteomic changes. From a pathophysiologic perspective, *H. pylori* manifests virulence by inducing chronic inflammation and releasing major cytotoxins (most notably CagA and VacA). Specifically, when *cag*PAI-carrying *H. pylori* attaches to a host cell, the type IV secretion system can translocate bacterial effector molecules into the host cell cytoplasm[33], which triggers a cascade of intracellular signaling events, leading to profound activation of pro-inflammatory cytokines, such as IL-1, IL-6, IL-8, TNF-a, NF-κB, and IFN-γ[2,34,35]. The current study corroborated these findings, revealing an upregulation of pathways related to immune activation, including phagocytosis, leukocyte activation, differentiation, and cell-cell adhesion, due to *H. pylori* infection. Furthermore, disturbances in protein folding and maturation, cell shape regulation, ATP biosynthetic process, and cellular response to hypoxia suggested a disturbed metabolic state.

H. pylori infection status was assessed in the Linqu set using the ELISA serologic antibody test and the CUBT for a

sub-group. The ELISA test indicates past and current infections and is influenced by factors like antibody half-life and *H. pylori* subtypes, whereas the CUBT primarily reflects the intensity of a current infection[36]. Nevertheless, despite these methodologic differences, the key proteins associated with *H. pylori* infection that were identified exhibited high consistency and were subsequently used effectively in the analysis.

Through comparative and co-analysis of proteomic profiles altered by *H. pylori* infection and the progression of gastric lesions, a concordance between these two patterns of change was detected. Most proteins that were upregulated in the *H. pylori*-positive group also exhibited an elevated expression trend in GCs. The phenomenon observed suggested a shared biological alteration that takes place throughout *H. pylori* infection and GC development by ruling out the grouping bias and covariates. Nevertheless, an asymmetric dispersion pattern of many proteins relative to the correlation curve for *H. pylori* infection and GC, particularly those proteins that were downregulated, may partially reflect the etiologic complexity underlying *H. pylori* pathogenicity and GC carcinogenesis even though *H. pylori* infection is recognized as a major risk factor for GC. Furthermore, although gastric lesions progress through a multi-stage cascade during gastric carcinogenesis, *H. pylori* infection is primarily implicated in the early stages of carcinogenesis. This early-stage involvement may also partly account for the heterogeneity observed in the differentially expressed proteins associated with *H. pylori* infection and GC.

scRNA analyses were also used to quantify the expression of key marker genes to delve deeper into the cellular level. By focusing on the epithelium, the findings unveiled an altered transcriptional landscape across the progression of cell groups representing healthy gastric mucosa (gastric epithelial cells-to-precancerous lesions) and IM cells, then to GC (malignant cells).

The upregulated markers shared by *H. pylori* infection and GC are consistent with several hallmark changes of carcinogenesis. First, the proliferative nature of malignant cells, which requires a substantial increase in protein synthesis and adequate energy production, consequently leads to the upregulation of pathways related to protein folding, maturation, and mRNA translation in these cells[37]. Specifically, *HNRNPM*, *PARP14*, *PARP1*, *TPP1*, and *TYMP* are involved in RNA splicing, ADP-ribosylation, telomerase activity, and thymine phosphorylation-associated DNA repair mechanisms[38,39]. The aberrant expression of these 5 genes suggests abnormal regulation in cell proliferation processes. Second, cancerous cells depend heavily on anaerobic respiration, a phenomenon known as the "Warburg effect," which accelerates glucose utilization and glycolysis. This dependency can also influence cellular metabolism and is attributed to alterations in membrane proteins and impaired lysosome function[40]. Upregulation of *ENO1*, *PLCG2*, and *GLS* (genes encoding enzymes vital for glycolysis, diacylglycerol and phosphorylcholine generation, and glutamine metabolism[41,42]) indicates metabolic reprogramming. Third, immune responses are also altered in gastric carcinogenesis. Upregulation of *PYCARD* suggests inflammasome formation, innate immune activation, and apoptosis[43], which may contribute to the inflammatory microenvironment characteristic of GC. Finally, *OLFM4*, a crucial stemness marker, has been reported to interact with *MYH9* to promote IM stem cell formation by activating the β-catenin-related Wnt signaling pathway[44]. In the current study *H. pylori* infection was shown to upregulate markers, such as *OLFM4* and *ENO1*, which exhibit an increasing trend from normal gastric tissue-to-IM, then to malignancy, thereby identifying key molecular events driving gastric carcinogenesis *via H. pylori*.

In contrast, the downregulated markers due to *H. pylori* infection and in GC were shown to be strongly associated with epithelium maintenance. Notably, *MUC5AC*, *GKN1*, and *TFF2*, which encode proteins crucial for preserving epithelium integrity and the homeostasis of the mucus layer, displayed reduced expression[45-47]. Furthermore, the downregulation of *GSN*, *IGFBP2*, and *TAGLN2*, which are involved in epithelial growth regulation and cytoskeleton reconstruction[48-50], may compromise the ability of the gastric mucosa to heal and recover from damage. In addition, decreased expression of *ALDH3A1* suggests that *H. pylori* infection may weaken the antioxidation and detoxication capacity of gastric mucosa[51], potentially facilitating the carcinogenic process.

Although endoscopic examinations are effective in preventing GC[52], endoscopic examinations are invasive, costly, and heavily reliant on skilled endoscopists and pathologists. Because a one-size-fits-all endoscopic screening approach is not feasible in China, utilizing reliable biomarkers derived from prospective evidence could be a practical strategy to identify and target high-risk populations, thereby supporting GC prevention efforts. Herein, the potential mechanisms underlying *H. pylori*-induced gastric carcinogenesis were deciphered and identified protein markers that could signify the progression of gastric lesions and the development of GC at an early stage. These findings led to the development of two novel panels [a tissue protein panel to guide the selection of

populations for repeated screening and a plasma protein-based panel (IGFBP2, GSN, ENO1, and OLFM4)] for risk stratification and early detection of GC. Integrating these proteomic signatures with multi-omics data and wet-lab exploration will provide mechanistic evidence on GC etiology and aid in the precision management of *H. pylori*-associated GC.

Several limitations to the current study should be acknowledged. First, the sample size was modest with prospective endoscopic follow-up based on the Linqu cohort, which may have constrained the power to explore tissue protein signatures for GC development. Nevertheless, the tissue protein risk score was significantly associated with the progression to neoplasia or GC as an alternative outcome, supporting the potential significance of the identified proteins. Second, a dataset that not only includes histopathologic diagnoses and *H. pylori* infection but also covers both gastric tissue and plasma proteomics is not available. Therefore, gastric tissue and plasma proteins could not be assessed in the same population. Third, information on *H. pylori* infection was not available for the Beijing cohort, the scRNA-seq databsase, and the UKB. Fourth, the proteomics platform tested in the UKB only encompassed several of the key tissue markers. To investigate other plasma proteins not included in the UKB panel, additional large prospective studies will be required in future research. Fifth, a limitation inherent to bulk-tissue proteomics analysis is the inability to fully capture the heterogeneity of the tumor microenvironment (TME). The complex intercellular interactions between epithelial compartments and stromal components remain incompletely characterized for the identified key epithelial protein markers. Future investigations employing single-cell proteomics, spatial transcriptomics, or multiplex immunohistochemistry techniques may be warranted to dissect the cell-type-specific protein signatures and spatially resolve molecular interactions within the TME during gastric carcinogenesis.

Conclusions

In summary, leveraging in-house and publicly available proteomics and scRNA-seq data, the current study utilized a mixed-design integrating case-control studies of proteomic signatures for *H. pylori* infection and risk of gastric lesions, exploration of biomarkers at the single-cell transcriptome level, and prospective studies on key proteins associated with the risk of gastric lesion progression and GC development, and unveiled key proteomics signatures associated with *H. pylori*-related gastric carcinogenesis. A panel of 15 tissue proteins was identified, demonstrating the potential for targeting individuals with a high risk of progressing to gastric neoplasia. Furthermore, a four-circulating protein panel may be used as non-invasive markers predicting the risk of GC development. While the current GC prevention programs rely mainly on gastroendoscopy (a method that is invasive, costly, and heavily skill-dependent) the current study developed novel non-invasive circulating protein biomarkers for *H. pylori*-induced gastric carcinogenesis. These biomarkers hold significant translational potential, enabling the identification of high-risk GC populations and facilitating improvements in targeted GC prevention strategies. Additional efforts are warranted to elucidate the mechanisms underlying key tissue proteins involved in *H. pylori*-induced gastric carcinogenesis and to validate the potential use as biomarkers, both in tissue and plasma levels, through multi-center prospective studies.

Acknowledgments

We are indebted and thankful to all participants for their valuable contributions. The authors would also like to thank the GEO database and UK Biobank Pharma Proteomics Project (UKB-PPP) consortium for providing their platforms and contributors for their meaningful datasets. This research was conducted using the UK Biobank Resource under Application No. 90999.

Grant support

This study was funded by the National Natural Science Foundation of China (Grant No. 82273704), Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant Nos. 2023ZD0501400-2023ZD0501402), Beijing Hospitals Authority's Ascent Plan (Grant No. DFL20241102), Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support (Grant No. ZLRK202325), Peking University Medicine Fund for World's Leading Discipline or Discipline Cluster Development (Grant No. BMU2022XKQ004), and the Science Foundation of Peking University Cancer Hospital (Grant Nos. BJCH2024BJ02, XKFZ2410, and 2022-27).

Conflict of interest statement

No potential conflicts of interest are disclosed.

Author contributions

Conceived and designed the analysis: Wenqing Li.
Adapted algorithms and software for data analyses: Yu Jin, Xue Li, Lanxin Yang, and Hengmin Xu.
Contributed to subject recruitment and sample collection: Yu Jin, Xue Li, Lanxin Yang, Zongchao Liu, and Yang Zhang.
Performed *H. pylori* antibody assays: Xue Li and Bingyao Cai.
Wrote and revised the manuscript: Yu Jin, Wenjing Zhao, Bingyao Cai, Wenqing Li and Kaifeng Pan.
All authors read and approved the submitted version.

Data availability statement

Our in-house proteomic datasets are deposited in the ProteomeXchange Consortium *via* the iProX partner repository (access No. IPX0003438002, DOI: 10.1016/j.ebiom.2021.103714). The scRNA dataset is available at the GEO database (access no. GSE249874, DOI: 10.1136/gutjnl-2023-iddf.110). The UKB-PPP proteomic dataset could be accessed by contacting the UKB-PPP Consortium at biobank.ndph.ox.ac.uk. Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon reasonable request.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024; 74: 229-63.
2. Correa P, Piazuelo MB. The gastric precancerous cascade. J Dig Dis. 2012; 13: 2-9.
3. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol. 2012; 13: 607-15.
4. Pan KF, Li WQ, Zhang L, Liu WD, Ma JL, Zhang Y, et al. Gastric cancer prevention by community eradication of *Helicobacter pylori*: a cluster-randomized controlled trial. Nat Med. 2024; 30: 3250-60.
5. Ford AC, Yuan Y, Moayyedi P. *Helicobacter pylori* eradication therapy to prevent gastric cancer: systematic review and meta-analysis. Gut. 2020; 69: 2113-21.
6. 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. Br Med J. 2019; 366: l5016.
7. Malfertheiner P, Camargo MC, El-Omar E, Liou JM, Peek R, Schulz C, et al. *Helicobacter pylori* infection. Nat Rev Dis Primers. 2023; 9: 19.
8. Koch MRA, Gong R, Friedrich V, Engelsberger V, Kretschmer L, Wanisch A, et al. CagA-specific gastric CD8(+)tissue-resident T cells control *Helicobacter pylori* during the early infection Phase. Gastroenterology. 2023; 164: 550-66.
9. Huang S, Guo Y, Li ZW, Shui G, Tian H, Li BW, et al. Identification and validation of plasma metabolomic signatures in precancerous gastric lesions that progress to cancer. JAMA Netw Open. 2021; 4: e2114186.
10. Liu ZC, Wu WH, Huang S, Li ZW, Li X, Shui GH, et al. Plasma lipids signify the progression of precancerous gastric lesions to gastric cancer: a prospective targeted lipidomics study. Theranostics. 2022; 12: 4671-83.
11. Wang R, Song S, Qin J, Yoshimura K, Peng F, Chu Y, et al. Evolution of immune and stromal cell states and ecotypes during gastric adenocarcinoma progression. Cancer Cell. 2023; 41: 1407-26.e9.
12. Guo Y, Zhang Y, Gerhard M, Gao JJ, Mejias-Luque R, Zhang L, et al. Effect of *Helicobacter pylori* on gastrointestinal microbiota: a population-based study in Linqu, a high-risk area of gastric cancer. Gut. 2020; 69: 1598-607.
13. Li X, Zheng NR, Wang LH, Li ZW, Liu ZC, Fan H, et al. Proteomic profiling identifies signatures associated with progression of precancerous gastric lesions and risk of early gastric cancer. EBioMedicine. 2021; 74: 103714.
14. Li N, Zhu Y, Liu J, Xu X, Zheng P, Lei Y, et al. IDDF2023-ABS-0273 single-cell transcriptomic profiling reveals molecular heterogeneity of fibroblast in *Helicobacter pylori*-associated gastric carcinogenesis. Gut. 2023; 72(Suppl 1): A125-7.
15. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023; 622: 329-38.
16. 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: 409-12.
17. Wik L, Nordberg N, Broberg J, Björkesten J, Assarsson E, Henriksson S, et al. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. Mol Cell Proteomics. 2021; 20: 100168.
18. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8: 14049.
19. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024; 42: 293-304.
20. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007; 18: 805-35.
21. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021; 2: 100141.

22. Džeroski S, Ženko B. Is combining classifiers with stacking better than selecting the best one? Mach Learn. 2004; 54: 255-73.
23. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019; 14: e0224365.
24. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019; 1: 206-15.
25. Charilaou P, Battat R. Machine learning models and over-fitting considerations. World J Gastroenterol. 2022; 28: 605-7.
26. Kumar V, Ramnarayanan K, Sundar R, Padmanabhan N, Srivastava S, Koiwa M, et al. Single-cell atlas of lineage states, tumor microenvironment, and subtype-specific expression programs in gastric cancer. Cancer Discov. 2022; 12: 670-91.
27. 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: 1934-47.e5.
28. Zhang M, Hu S, Min M, Ni Y, Lu Z, Sun X, et al. Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing. Gut. 2021; 70: 464-75.
29. Jiang S, Qian Q, Zhu T, Zong W, Shang Y, Jin T, et al. Cell taxonomy: a curated repository of cell types with multifaceted characterization. Nucleic Acids Res. 2023; 51(D1): D853-60.
30. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single cell portal: an interactive home for single-cell genomics data. bioRxiv. 2023.
31. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford). 2019; 2019: baz046.
32. Lunardon N, Menardi G, Torelli N. ROSE: a package for binary imbalanced learning. R J. 2014; 6: 82-92.
33. Zavros Y, Merchant JL. The immune microenvironment in gastric adenocarcinoma. Nat Rev Gastroenterol Hepatol. 2022; 19: 451-67.
34. Mejías-Luque R, Zöller J, Anderl F, Loew-Gil E, Vieth M, Adler T, et al. Lymphotoxin β receptor signalling executes *Helicobacter pylori*-driven gastric inflammation in a T4SS-dependent manner. Gut. 2017; 66: 1369-81.
35. Li X, Pan K, Vieth M, Gerhard M, Li W, Mejías-Luque R. JAK-STAT1 signaling pathway is an early response to *Helicobacter pylori* infection and contributes to immune escape and gastric carcinogenesis. Int J Mol Sci. 2022; 23: 4147.
36. Li ZX, Bronny K, Formichella L, Mejías-Luque R, Burrell T, Macke L, et al. A multiserological line assay to potentially discriminate current from past *Helicobacter pylori* infection. Clin Microbiol Infect. 2024; 30: 114-21.
37. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022; 12: 31-46.
38. Wang X, Li J, Bian X, Wu C, Hua J, Chang S, et al. CircURI1 interacts with hnRNPM to inhibit metastasis by modulating alternative splicing in gastric cancer. Proc Natl Acad Sci U S A. 2021; 118: e2012881118.
39. Aguiar RC, Takeyama K, He C, Kreinbrink K, Shipp MA. B-aggressive lymphoma family proteins have unique domains that modulate transcription and exhibit poly(ADP-ribose) polymerase activity. J Biol Chem. 2005; 280: 33756-65.
40. Eriksson I, Öllinger K. Lysosomes in cancer-at the crossroad of good and evil. Cells. 2024; 13: 459.
41. Huang H, Tang S, Ji M, Tang Z, Shimada M, Liu X, et al. p300-mediated lysine 2-hydroxyisobutyrylation regulates glycolysis. Mol Cell. 2018; 70: 663-78.e6.
42. Rumping L, Tessadori F, Pouwels PJW, Vringer E, Wijnen JP, Bhogal AA, et al. GLS hyperactivity causes glutamate excess, infantile cataract and profound developmental delay. Hum Mol Genet. 2019; 28: 96-104.
43. Zhou R, Yazdi AS, Menu P, Tschopp J. A role for mitochondria in NLRP3 inflammasome activation. Nature. 2011; 469: 221-5.
44. Wei H, Li W, Zeng L, Ding N, Li K, Yu H, et al. OLFM4 promotes the progression of intestinal metaplasia through activation of the MYH9/GSK3β/β-catenin pathway. Mol Cancer. 2024; 23: 124.
45. Krishn SR, Ganguly K, Kaur S, Batra SK. Ramifications of secreted mucin MUC5AC in malignant journey: a holistic view. Carcinogenesis. 2018; 39: 633-51.
46. Xing R, Cui JT, Xia N, Lu YY. GKN1 inhibits cell invasion in gastric cancer by inactivating the NF-kappaB pathway. Discov Med. 2015; 19: 65-71.
47. Kjellev S. The trefoil factor family - small peptides with multiple functionalities. Cell Mol Life Sci. 2009; 66: 1350-69.
48. Nag S, Ma Q, Wang H, Chumnarnsilpa S, Lee WL, Larsson M, et al. Ca2+ binding by domain 2 plays a critical role in the activation and stabilization of gelsolin. Proc Natl Acad Sci U S A. 2009; 106: 13713-8.
49. Li T, Forbes ME, Fuller GN, Li J, Yang X, Zhang W. IGFBP2: integrative hub of developmental and oncogenic signaling network. Oncogene. 2020; 39: 2243-57.
50. Jin H, Zheng W, Hou J, Peng H, Zhuo H. An essential NRP1-mediated role for Tagln2 in gastric cancer angiogenesis. Front Oncol. 2021; 11: 653246.
51. Lee JS, Kim SH, Lee S, Kang JH, Lee SH, Cheong JH, et al. Gastric cancer depends on aldehyde dehydrogenase 3A1 for fatty acid oxidation. Sci Rep. 2019; 9: 16313.
52. Li WQ, Qin XX, Li ZX, Wang LH, Liu ZC, Fan XH, et al. Beneficial effects of endoscopic screening on gastric cancer and optimal screening interval: a population-based study. Endoscopy. 2022; 54: 848-58.

Cite this article as: Jin Y, Li X, Cai B, Yang L, Zhao W, Xu H, et al. Proteomic profiling and scRNA sequencing identify signatures associated with *Helicobacter pylori* infection and risk of developing gastric cancer. Cancer Biol Med. 2025; 22: 946-963. doi: 10.20892/j.issn.2095-3941.2025.0077