Plasma protein biomarkers for the early detection of gastric preneoplasia and cancer: a prospective study

Quentin Giai Gianetto[1,2], Valérie Michel[3], Thibaut Douché[1], Karine Nozeret[3], Aziz Zaanan[4], Oriane Colussi[4a], Isabelle Trouilloud[4b], Simon Pernot[4c], Marie-Noelle Ungeheuer[6], Catherine Julié[7], Nathalie Jolly[8], Julien Taïeb[5], Dominique Lamarque[9d], Mariette Matondo[1], Eliette Touati[3*]

[1]Institut Pasteur, Proteomics Core Facility, MSBio, UAR CNRS 2024, Université Paris Cité, Paris, France
[2]Institut Pasteur, Bioinformatics and Biostatistics Hub, Université Paris Cité, Paris, France
[3]Institut Pasteur, Infection, Genotoxicity and Cancer Team, CNRS UMR 6047, Université Paris Cité, Paris, France
[4]Department of Gastroenterology and Digestive Oncology, Georges Pompidou European Hospital, Assitance Publique-Hôpitaux de Paris, Université Paris-Cité, Paris, France
[5]Department of Gastroenterology and Digestive Oncology, Georges Pompidou European Hospital, SIRIC CARPEM, Université Paris-Cité, Paris, France
[6]Institut Pasteur, Clinical Investigation and Access to Clinical Resources (ICAReB), Center for Translational Research, Paris, France.
[7] Pathology Unit Hôpital Ambroise Paré, Assitance Publique-Hôpitaux de Paris,CHU Paris IdF Ouest, Boulogne-Billancourt, France
[8]Institut Pasteur, Clinical Research Coordination Office, Université Paris Cité, Paris, France
[9]Unit of Hepato-Gastroentroenterology, Hôpital Ambroise Paré, Assitance Publique-Hôpitaux de Paris CHU Paris IdF Ouest, Boulogne-Billancourt, France

Running title: **Protein markers of early gastric cancer**

Present address: [a]Gastroenterology and Hepatology Unit, Antony Private hospital, Antony, France; [b]Departement of Oncology, AP-HP, Hôpital St Antoine, Paris, France; [c]Institut Bergonié, Bordeaux, France ; [d] Reference center for Anorectal and pelvic malformations, University Hospital Necker-Enfants malades, Department of Pediatyric Visceral Surgery, Paris, France

*Corresponding author: eliette.touati@pasteur.fr

Abstract

Background: Gastric cancer (GC) is often associated with a poor prognosis, due to its asymptomatic early stages. The current gold standard for diagnosing GC, upper endoscopy, is invasive and has limited sensitivity for detecting gastric preneoplasia such as dysplasia. Non-invasive biomarkers, such as circulating proteins in the blood, hold promise for the early detection of asymptomatic gastric lesions.

Methods: During an exploratory study, plasma samples from 39 participants, including patients diagnosed with different gastric pathologies (gastritis, preneoplasia, GC) and healthy controls, were analyzed using mass spectrometry-based proteomics. Data are available via ProteomeXchange with identifier PXD062315. Differentially abundant proteins were identified through pairwise comparisons and sparse Partial Least Squares Discriminant Analysis (sparse PLS-DA). Fifteen candidate proteins were selected and quantified by ELISA in plasma samples from the full cohort of 138 participants.

Results: Proteomic profiling identified 691 proteins in plasma samples. Pairwise comparisons highlighted up to 213 candidate biomarkers capable of distinguishing cancer patients from healthy controls, while distinguishing gastritis and preneoplasia proved challenging due to their similar proteomic profiles. Sparse PLS-DA identified 85 proteins that distinguish all patient groups. Subsequent ELISA validation confirmed Leptin as a promising biomarker for detecting gastric preneoplasia, particularly in women. Additionally, nine additional proteins (ATAD3B, IGFALS, JUP, LBP, MAN2A1, ARG1, CA2, HPT, KRT14) showed differential plasma levels across patients groups, with variations influenced by age and gender. Multi-biomarker prediction models incorporating factors such as age, gender, and *H. pylori* status, outperformed single-protein models. Using 4-fold cross-validation repeated 10 times, the top-performing models achieved high predictive accuracy, with mean AUROC values of 85.3% for classifications cancer

vs. non-cancer and 83.9% for cancer/preneoplasia *vs*. healthy/gastritis. On the full cohort, optimal biomarker combinations reached AUROC values exceeding 94% for these classifications.

Conclusions: This study introduces a novel, non-invasive approach for predicting GC based on plasma protein concentrations across a broad spectrum of gastric pathologies. Predictive models, validated through robust cross-validation, demonstrated high accuracy using a limited panel of biomarkers. Leveraging simple blood sampling, this strategy holds promise for high-risk gastric mucosal lesions, including at asymptomatic stages. Such an approach could significantly improve early detection and clinical management of GC, offering direct benefit for patient outcomes.

Keywords: Intestinal metaplasia, Dysplasia; Gastric cancer; Liquid biopsy; Blood; Diagnostic protein biomarkers; Mass spectrometry-based proteomics; LC-MS/MS

Background

Gastric cancer (GC) is the fifth most commonly diagnosed cancer and the fourth leading cause of cancer-related death worldwide (1). With over 1 million new cases reported annually, GC remains an important healthcare challenge. It predominantly affects individuals over the age of 60, preferentially men, and it is strongly associated with *H. pylori* infection (2). *H. pylori* infects nearly half of the global population and is recognized as a major risk factor for GC (3). While the global incidence and mortality of GC have declined over the past decades, a concerning rise cases among youngers (< 50 years) has been observed in both low- and high-income countries (4) (5). Due to its asymptomatic phenotype during the first steps of its development, GC is mainly associated with a poor prognosis, highlighting the importance of its early detection. The incidence of GC is higher in Eastern Asia where national screening programs have been implemented as Japan and South Korea, leading to a 5-year survival of about 60%, compared to 15-20% in Western countries, thus highlighting the benefit of its prevention (6).

The majority of GC cases are adenocarcinoma that develop through a multistep process, initiated by a chronic inflammation and progressing through atrophic gastritis (AG), intestinal metaplasia (IM) and dysplasia (DYS), ultimately leading to cancer lesions (7). If detected at an early stage, meaning AG and/or low-grade preneoplasia, GC can be a preventable disease (8). The prevalence of *H. pylori*-associated AG and IM, increases logarithmically with age and confers higher risk of GC (9). While *H. pylori* eradication can reduce the risk of GC, residual risk persists, especially in individuals with existing gastric preneoplasia or high baseline risk (10) (11). Presently, upper endoscopy is the gold standard for GC diagnosis and preneoplasia monitoring, as recommended by the European Society for Gastrointestinal Endoscopy (12) (13). However, this method is invasive, costly and often limited in detecting transient lesions such as DYS (12) (13). Recent studies have reported that 10-20% of patients of GC cases had no preneoplastic lesions detected by endoscopy within 6 to 36 months prior to the diagnosis (14).

Therefore, the discovery of non-invasive biomarkers based on liquid biopsies as blood is of paramount interest to identify the presence of gastric lesions at an asymptomatic stage. It is crucial to reduce GC global burden, not only for the screening and early detection of GC but also for patient follow-up, to anticipate disease recurrence and treatment monitoring.

Circulating proteins have emerged as promising biomarkers for cancer detection, as previously reported for the CancerSeek test (15). While glycoproteins such as carbohydrate antigen 12-5 (CA12-5), 19-9 (CA19-9) and carcinoembryonic antigen (CEA) are currently used in clinical practice, their sensitivity and specificity for GC remain limited (16)(17). Similarly, serological markers like pepsinogen I and II (PGI and PGII) and PGI/PGII ratio show some predictive value for AG- particulary in gastric corpus- with a sensitivity ranging from 67% to 85% and a specificity from 73% to 87% (18)(19). Despite many efforts, none of the reported blood-based GC biomarkers have been translated into clinical practice to date. Additionally, very few studies have specifically investigated blood biomarkers associated with DYS and early GC lesions (20). Recently, developments and improvements in high-throughput technologies, particularly mass spectrometry-based proteomics, have enabled the identification of biomarker panels with improved diagnostic accuracy compared to single proteins. For instance, a 19-proteins signature has been identified to distinguish early-stage (I/II) GC from healthy subjects (21). In our study, we identified differential plasma proteins associated with gastric preneoplasia and GC using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) analysis. We then validated some of these findings using an orthogonal technique, ELISA, which confirmed their potential as biomarkers. Our study reports promising candidates for biomarker panels that could enhance early detection and risk stratification for individuals predisposed to GC, thereby supporting more effective prevention strategies.

Methods

Study population

The study cohort consists of 138 participants. It includes 32 healthy asymptomatic volunteers recruited at the Institut Pasteur clinical investigation and biomedical research support unit (ICAReB-Clin). They were confirmed for their *H. pylori*-negative serology using a commercial Enzyme-Linked Immunosorbent Assay (ELISA), Serion ELISA Classic (Serion/diagnostic, Denmark). The groups of patients included 26 cases of non-AG referred as "gastritis", 20 of AG/preneoplasia, referred as the "preneoplasia" group, including patients with AG either with IM or DYS or not (Table 1). Patients with gastritis and preneoplasia were recruited at the unit of Hepato-Gastroenterology, AP-HP, A. Paré hospital, Boulogne-Billancourt, France. The 60 patients with GC (stages III and IV) were enrolled from the unit of Digestive Oncology, AP-HP, European Hospital G. Pompidou (HEGP), Paris, France. All patients were previously diagnosed through upper endoscopy combined with gastric biopsies and anatomopathology analysis. All patients were adults, they had not received antibiotherapy and had not been treated with bismuth compounds, proton pump inhibitors or non-steroidal anti-inflammatory drugs for at least two weeks prior to the study. Each patient was informed about the study and provided written consent. For each patient, 10 ml of blood were collected and transferred to Leucosep® tubes (Dutscher, France) for plasma isolation according to recommendation of the supplier.

The study was approved by the Institut Pasteur Clinical Research Coordination Office (Ref 2013-29), the French advisory committee on information processing in Health Research (Ref 14.202) and the French data protection authority: Commission Nationale de l'Informatique et des Libertés (CNIL) (Ref: 2017-204).

Identification and quantification of plasma proteomes using LC-MS/MS

Plasma samples for LC-MS/MS analysis. As indicated in Table 1, plasma proteome identification

and quantification were performed on samples from a subset of cohort patients with gastritis (n=10), preneoplasia (n=9), GC (n=10), and healthy controls (n=10). The most abundant proteins were first depleted from plasma, and the remaining protein samples underwent proteolytic digestion, as detailed in the Supplementary material. Next, proteins identification and quantification were performed using Data-Independent Acquisition (DIA), based on a spectral library obtained through Data-Dependent Acquisition (DDA).

LC-MS/MS analysis. Mass spectrometry-based plasma proteomics were conducted in two steps. The 1st step was dedicated to the spectral library generation from DDA using the fractionated "pool" sample. The 2nd step was defined to specifically identify and quantify plasma proteome for each patient group from DIA. For this purpose, a nanochromatographic system (Proxeon EASY-nLC 1200 - Thermo Fisher Scientific, Waltham, Massachusetts, USA) was coupled on-line to a Q Exactive™ HF Mass Spectrometer (Thermo Fisher Scientific) using an integrated column oven (PRSO-V1 - Sonation GmbH, Biberach, Germany). For each sample, 1 µg of peptides was injected onto a home-made C18 column. Peptides elution and mass spectra acquisition were performed as described in the Supplementary material.

Data processing for proteins identification and quantification including building of the spectral library and data analysis for DIA method are reported in the Supplementary material. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (22) with the dataset identifier PXD062318.

Statistical analysis of large-scale LC-MS/MS proteomic data for biomarker identification

Multivariate analysis of the MS-based proteomics data. Pairwise correlation analysis, hierarchical clustering and partial least square-discriminant analysis (PLS-DA) have been performed to highlight global similarities between quantified proteomes of plasma samples. The correlation matrix represents the Pearson correlation coefficients between each pair of samples

computed using all complete pairs of intensity values measured in these samples. The hierarchical cluster analysis has been conducted via multiscale boostrap resampling (1000 bootstrap replications) with the Ward's method and a correlation-based distance measure thanks to the *pvclust* function of the *pvclust* R package (23), after log2 transformation of the intensities, imputation of the missing values with the *impute.slsa* function of the *imp4p* R package (24) and a normalization using a sample-median centering method inside conditions. PLS-DA using 5 components was performed using the *mixOmics* R package (25) to assess whether the quantified proteomes differ among the four patient groups (H, Gastritis, Preneoplasia, and GC).

Identification of potential biomarkers from MS-based proteomics data. To identify potential protein biomarkers, pairwise differential analyses (e.g. comparisons of one category of patients *vs* another) were conducted, as well as a multivariate sparse PLS-DA.

To identify proteins that were more abundant in one condition than another, quantified intensities obtained using Spectronaut were compared. Proteins were retained for further statistical analysis only if they had at least 2 quantified intensity values (for the "relaxed" analysis) or 9 quantified intensity values (for the "strict" analysis) in at least one of the two conditions being compared. Proteins that were quantified in one condition but not in the other were set aside, as they could be directly considered as potential biomarkers. Following this initial filtering, the intensities of the remaining proteins were log-transformed (log2). Next, intensity values were normalized by median centering within a same category of patients (section 3.5 in (26)). Missing values were imputed using the *impute.slsa* function from the *imp4p* R package (24). Proteins with a fold-change (FC) below 2 (i.e., absolute $\log_2(FC) < 1$) were considered not significantly differentially abundant. Statistical testing was then performed using a *limma* t-test, implemented in the *limma* R package (27). To control for multiple testing, an adaptive Benjamini-Hochberg procedure was applied to the resulting p-values using the *adjust.p* function from the *cp4p* R package (28). The robust method described in (29) was used to estimate the proportion of true null hypotheses

among the statistical tests. Proteins with an adjusted p-value below a false discovery rate (FDR) of 1% were considered significantly differentially abundant. Ultimately, proteins identified as potential biomarkers were those that emerged from this multiple testing analysis, along with those quantified in one condition but not in the other. The lists of potential biomarkers obtained from the different comparisons were next compared using Venn diagrams to highlight proteins that are differentially abundant in several comparisons.

A sparse PLS-DA was performed alongside pairwise comparisons of patient categories. Unlike the previous approach, which compares patient groups two by two, sparse PLS-DA aims to distinguish all groups simultaneously using a restricted set of proteins. For this analysis, only proteins with at least two quantified values in all patient categories were considered. Similar to the differential analyses, the intensities of the selected proteins were log-transformed (log2), normalized by median centering within a same category of patients (section 3.5 in (26)) , and missing values were imputed using the *impute.slsa* function from the *imp4p* R package (24). Finally, sparse PLS-DA was applied, with a constraint of selecting only 10 proteins per component, using the *mixOmics* R package (25).

Validation of LC-MS/MS-identified proteins using Enzyme Linked Immunosorbent Assays (ELISA), statistical analysis and prediction models

To validate the protein biomarker candidates deduced from the analysis of the LC-MS/MS data, the plasma concentration of 15 short-listed proteins were measured across the full cohort of patient plasma samples (138 samples) using ELISA assays according to the manufacturer's instructions (MyBiosource, USA and R&D systems, USA) (Supplementary material- Table S1). Statistical analysis of the data was performed using the Student's t-test or Mann-Whitney test. Results were considered significant if $p<0.05$.

Correlation analysis of measured concentrations. The measured concentrations were analyzed

using Pearson correlation coefficients between biomarkers, using all complete pairs of concentration values measured in these samples. A hierarchical cluster analysis has been conducted with the Ward's method and a correlation-based distance measure derived from the correlation coefficients. This approach was used to identify biomarkers displaying similar concentration patterns across patients.

Estimation of prediction models to distinguish patient categories from measured concentrations. To estimate prediction models for the different pathologies, we used a 4-fold cross validation strategy repeated 10 times. It consists of randomly splitting the set of patients into four sets while preserving the proportion of patients associated to the different pathologies in each as in the original dataset. Three sets of patients over four are then used alternatively to create a training dataset used to estimate the prediction model, while the fourth set is used to assess the performance of the estimated model (test dataset). An up-sampling technique was used to duplicate patients of the minority categories with replacement to tackle the imbalance of the categories in the training dataset in view to improving model performance. Finally, each model provides an estimated probability that a patient is affected by a pathology from a set of biomarkers. We assessed the results of the estimated prediction models on the remaining set of patients (test dataset) using Area Under the ROC curve (AUROC). For each model, we randomly repeated 10 times the 4-fold cross validation strategy to assess the variations of results. AUROC values and ROC curves have been obtained using the pROC R package (30). Confidence intervals have been obtained with the DeLong methods (31).

Different types of prediction models were estimated. Firstly, we estimated a model predicting all categories of patients (i.e. predicting whether the patient is suffering from GC, preneoplasia, gastritis or if he/she is healthy). Second, we estimated four binary models specific to each category (i.e. predicting healthy *vs* non-healthy, cancer *vs* non-cancer, etc.). Third, we evaluated a model grouping GC and preneoplasia in one category and healthy and gastritis in another (i.e.

predicting GC/preneoplasia *vs* healthy/gastritis). This approach results in six types of prediction models, each of them is estimated by both linear logistic regression and using random forests, leading to twelve estimated models for each combination of biomarkers.

Results

Characteristics of the studied cohort and patient diagnosis

The general characteristics of the patients included in the cohort are reported in Table 1. 58% and 61% of patients with gastritis or preneoplasia respectively, were *H. pylori*-positive compared to 28% in the GC group. All healthy individuals in the control group are *H. pylori*-negative. Mean age for the healthy group was significantly lower (42 years), compared to gastritis, preneoplasia and GC patient groups (59, 70 and 62 years; $P=0.0003$, $P<0.0001$ and $P<0.0001$, respectively). All patients were previously diagnosed with upper gastrointestinal (GI) endoscopy, and those with AG had undergone upper GI endoscopy within 12 months prior to inclusion. Endoscopic examination of the mucosa was performed under white and blue light, in order to obtain biopsies from sites with the most severe lesions of atrophy or metaplasia, in accordance with European recommendations (12). In addition, 18 biopsy samples were systematically taken from the four sides of the antrum, angulus and the four sides of the gastric body.

Multivariate analysis of the LC-MS/MS-based proteomics data

An initial exploratory phase was conducted using LC-MS/MS-based proteomics on plasma from 39 individuals, including healthy controls and patients diagnosed with gastritis, preneoplasia and GC (Table 1). Following plasma protein preparation (see Supplemenatry material), a total of 691 proteins were identified and quantified.

To assess the similarity of proteomic profiles across groups, correlation analysis was performed (Figure 1). Overall, the proteomes showed high levels of correlation, with a minimum of 80.05%.

Within-group correlation was particularly strong, especially among patients within the same diagnostic category. Notably, proteomes from gastritis and preneoplasia patients were closely aligned, whereas those from healthy individuals and GC patients are more distinct from other patient groups. PLS-DA was used to further explore group separation based on proteomic profiles. The first two PLS-DA components clearly distinguished healthy individuals from patients affected by a pathology. Components 2 and 3 allow distinguishing GC patients from others (Figure 1A). As observed from the correlation analysis (Figure 1B), the separation between patients with gastritis from those with preneoplasia is low but still occurs at the level of the fourth component of the PLS-DA that explains only 3% of the total variance (Figures 1A). Together, the first four PLS-DA components allowed for the discrimination of the different patient categories based on their proteomes. These findings indicate distinct proteomic signatures in plasma samples associated with each group of patients.

Biomarker candidates identified by LC-MS/MS-based data.

To identify candidate biomarkers capable of distinguishing between patient groups, we performed differential analyses using multiple moderated t-tests for all pairwise comparisons ("*Two by two"* comparisons), as well as a sparse PLS-DA analysis, as described in the Methods section.

"Two by two" comparisons. For comparison of one pathology *vs* another, two types of statistical analyses were conducted: a "relaxed" analysis by imposing, for each protein, at least two quantified values in the plasma samples of patients from one of the two compared categories (Figure 2A), and a "strict" analysis by imposing at least nine quantified values in one of the two compared categories (Figure 2B). Since these proteins are quantified in almost all patients in a category, we can expect the "strict" analysis to give more robust results. Six different comparisons were performed: each group of patients *vs* healthy, gastritis *vs* preneoplasia, and

gastritis or preneoplasia *vs* GC. From the pairwise comparisons resulting from the "relaxed" analysis (Figure 2A), 408 proteins were found significantly more abundant in one category of patients than in another, among which 309 with differential level in at least two comparisons (see additional file 1 and Figure S1). With the "strict" analysis approach (Figure 2B), 236 proteins show differential level comparing two groups: 149 are differential in at least two comparisons (see data in Supplementary material Table S3 and additional file 1). It should be noted that the comparison of preneoplasia and gastritis categories show the fewest proteins differentially abundant, with 58 differential proteins found in the "relaxed" analysis (Figure 2A, Figure S1 and additional file 1), while none is found with the "strict" analysis (Figure 2B, Figure S2 and additional file 2). These 58 proteins are affected by missing values, which raises questions of their relevance. Considering the "strict" analysis, 2 proteins appear with differential level in 5 comparisons over 6: TFRC and IGHG1; 7 proteins in 4 comparisons over 6: PRCP, SAA1/2, PRSS3, MAN2A1, DCD, GPLD1 and IL1RAP. These proteins could serve as promising biomarker candidates for distinguishing between patient categories. In addition to pairwise condition comparisons, we performed a sparse PLS-DA analysis to identify a refined set of proteins capable of distinguishing all patient groups effectively.

*Sparse PLS-DA***.** Unlike the previous approach that compares each category of patients two by two, this sparse PLS-DA analysis attempts to separate all the groups according to a restricted number of proteins: we imposed the choice of only 10 biomarkers by component of the PLS-DA. The first component distinguishes healthy subjects from others, while the second one separates GC patients from those with gastritis and gastric preneoplasia (Figure 2C). The third component allows a separation between gastritis and preneoplasia. The correlation circle plots highlight the correlation of each protein selected by the sparse PLS-DA with its different components (Figure 2D). Among the 85 proteins identified as potential biomarkers by the sparse PLS-DA analysis, F13A1 is highly correlated with the first component, while TFRC is inversely correlated. They

could thus be biomarkers to separate healthy individuals from others. Other proteins like DCD, KRT19, KIF20B, ATAD3B, S100A12 or LEP are positively correlated with the second component, while MAN2A1, LBP or SVEP1 are inversely correlated. Thus, these proteins could contribute to separate GC patients from gastritis and preneoplasia patients. When examining proteins correlated with the third component separating gastritis and preneoplasia, CAPZA1, FUCA2, ERAP2 and LEP are found, also positively or negatively correlated with the second component separating GC patients from others. Interestingly, LEP is the only protein kept by the sparse PLS-DA with an absolute correlation close or superior to 0.5 with the three components. This protein was also found differentially abundant in the "relaxed" analysis in four comparisons (GC *vs* gastritis, GC *vs* healthy, preneoplasia *vs* healthy and GC *vs* preneoplasia) (additional file 2) and could thus be an interesting candidate, also considering its digestive function and its previously reported association with increased risk of IM and GC (32) (33). Importantly, most of the proteins highlighted by the sparse PLS-DA, are categorized as potential prognostic biomarkers of diverse cancers in the Human Protein Atlas (https://www.proteinatlas.org) (34), thus confirming their relevance in our study and their potency as biomarker candidates associated to gastric preneoplasia or GC.

Short-listing of 15 biomarker candidates from the LC-MS/MS analysis. Based on the analysis of LC-MS/MS data, 15 proteins were short-listed for further confirmation by ELISA, considering both the results of the statistical analyses, their biological functions and their relevance in cancer (Table 2). Statistically, all these proteins were found differentially abundant either in the "GC *vs* preneoplasia" or "GC *vs* heathy" comparison using the "strict" statistical analysis and could thus be biomarker candidates of GC. They were also significantly differentially abundant in at least one other comparison when using the "strict" statistical analysis, except LEP which presents some missing values that exclude it from this analysis. However, LEP was also particularly interesting to test by ELISA, according to the sparse PLS-DA results (Figure 2D). In addition, 7

of these 15 proteins were also differentially abundant between patients affected by preneoplasia and healthy controls in the "relaxed" statistical analysis (CA2, DCD, F13A1, HPT, IGFALS, KRT19, and MAN2A1) and could thus be potential biomarkers of gastric preneoplasia.

ELISA-based validation of 15 biomarker candidates highlighted by LC-MS/MS

ELISAs are the cornerstone of diagnostic testing in a wide range of applications due to their quantification capabilities, sensitivity and ability to quickly provide results. In this study, ELISAs were used to measure the plasma concentrations of selected biomarker candidates. The ELISA results confirmed LEP as a potent biomarker, in agreement with the findings from LC-MS/MS analyses. Its plasma level is differentially abundant when comparing patients with gastritis or preneoplasia *vs* healthy individuals, and patients with preneoplasia with those either with gastritis or GC (Figure 3). It is higher in the group of patients with preneoplasia than in the other categories and significantly lower in men than women for all groups, except GC patients (Figure S3). These findings suggest that plasma LEP concentration could serve as a robust biomarker for identifying preneoplastic lesions.

In addition to LEP, differential plasma levels were observed between patients with preneoplasia and healthy subjects for ATAD3B, IGFALS, JUP, LBP and MAN2A1 (Figure 3). These proteins, along with ARG1, CA2, HPT and KRT14 showed differential plasma level between GC patients and healthy individuals. Gender affects the plasma concentration of ATAD3B, JUP and MAN2A1 which showed differential levels specifically in women when comparing each group of patients to healthy individuals (Figure S3). When comparing GC patients to healthy individuals, differential plasma levels are observed in women for HPT and in men for IGFALS. Age can also have an impact. ARG1 is significantly more abundant in GC patients up to 50 years-old than in other groups (Figure S4). These 9 proteins could thus be considered as potent

biomarkers of gastric preneoplasia or GC. In some cases, their discriminating properties can be modulated by age or gender.

Four proteins, KRT19, F13A1, DCD, KIF20B, do not exhibit significant differences in plasma concentration across the patient categories (Figure 3). However, in younger patients, DCD levels display significant difference between GC and healthy groups (Figure S4) and are lower in men with GC compared to those with preneoplasia (Figure S3).

Although some of these proteins showed a promising trend, defining specific concentration thresholds for reliably diagnosing a given pathology in individual patients remains challenging. As reported in previous studies, single protein biomarkers often lack sufficient discriminatory power on their own. However, their diagnostic accuracy can be significantly enhanced when combined with other candidate biomarkers (20). In the next, we investigated the relationships between the biomarkers concentration and evaluated the added value of multi-marker combinations to improve patient classification.

Correlations between plasma concentration levels of the biomarkers.

Pearson correlation coefficients analysis between the biomarker concentrations in patients showed that many of them appear strongly significantly different from zero with a p-value threshold of 1% (Figure 4A). Many of the biomarkers are positively correlated. Their concentration evolves globally similarly in the plasma of patients, as KRT19, CA2 and LBP between them, HPT with DCD, DCD with CA2, JUP with LBP, KIF20B and ARG1 with F13A1, MAN2A1 with ATAD3B. Inversely, MAN2A1 is negatively correlated with IGFALS and CA2, as S100A12 and ATAD3B, or LEP with patient gender (i.e. higher levels in women than men). These observations led to four main protein clusters when using hierarchical clustering based on an absolute Pearson correlation distance (Figure 4B). The first cluster contains KRT19, S100A12, JUP, LBP, CA2, DCD and HPT, the second LEP and gender, the third includes age

and *H. pylori* infection status of the patient with ATAD3B, MAN2A1, IGFALS and KRT14, while the fourth is composed of F13A1, KIF20B and ARG1. These different behaviors of biomarkers therefore indicate that not all biomarkers provide the same information. The use of their combination can thus be a positive strategy leading to a more accurate classification of patients.

Performance of prediction models with single and multiple biomarkers

As described in the Methods section, six types of prediction models were developed using either linear logistic regression or random forests, evaluated through 4-fold cross-validation repeated 10 times (Figure 5A). Models were evaluated based on the mean and standard deviation of the AUROC criterion, as calculated from the test sets of the cross-validation process (Additional file 3). We first evaluated each of the 15 ELISA-validated protein biomarkers individually, also including age, gender and *H. pylori* infection status as additional factors that can be used in our prediction models to eventually improve their performance, thus resulting in 18 biomarkers.

Single biomarker analysis. We first assessed the predictive performance of individual protein biomarkers incorporated into the models. Figure 5B presents the mean AUROC (mAUROC) values for all biomarkers across different models, where the reported mAUROC corresponds to the highest value obtained between logistic regression and random forest models (see also Additional file 3). For models predicting all patient categories, the best-performing biomarker was the plasma concentration of KRT14 (mAUROC = 63.3%, sd = 4.4%), closely followed by patient age (mAUROC = 63.2%, sd = 4.8%). Notably, age emerged as a crucial factor, yielding the highest mAUROC for several classification tasks, including healthy *vs* non-healthy (mAUROC = 80.1%, sd = 6.9%), preneoplasia *vs* non-preneoplasia (mAUROC = 71.8%, sd = 8.8%), and GC or preneoplasia *vs* healthy or gastritis (mAUROC = 75.1%, sd = 8.1%). These results reflect the structure of our cohort with healthy individuals younger than in the other

categories of patients. For cancer vs non-cancer prediction, the best-performing biomarker was the plasma concentration of CA2 (mAUROC = 78.1%, sd = 8.7%). Meanwhile, for gastritis *vs* non-gastritis classification, the highest mAUROC was again obtained with KRT14 (mAUROC = 72.3%, sd = 8.4%). Although these results demonstrate the potential utility of single biomarkers, their moderate mAUROC values indicate that individual biomarkers alone may not be sufficient for robust disease classification.

Multiple biomarker analysis. To improve predictive accuracy, we examined models incorporating multiple biomarkers using the same methodology. Given that the number of possible biomarker combinations increases exponentially with the number of biomarkers included (Figure 5C), we limited our analysis to a maximum of six biomarkers selected from the 18 candidates, resulting in a total of 37,179 unique tested biomarker combinations (Additional file 3). As shown in Figure 5C, models that integrated multiple biomarkers outperformed those using a single one, achieving higher mAUROC values. However, adding more biomarkers did not always lead to better performance, as none of the highest mAUROC values were observed in models containing six biomarkers. For models predicting all patient categories, the best-performing model achieved a mAUROC of 77.9% (sd = 7.7%) and incorporated age, gender, *H. pylori* status and the plasma concentrations of ARG1 and KRT14 (Figure 5C). A different model, designed to predict cancer or preneoplasia *vs* healthy or gastritis, yielded an even higher mAUROC of 83.9% (sd = 5.1%) and was based on the plasma concentrations of ARG1, CA2, HPT, MAN2A1, and LBP. HPT also emerged as a key biomarker to distinguish between healthy and pathological states. When combining HPT with age and *H. pylori* status, this model achieved the highest mAUROC of 88.0% (sd = 5.3%) for this task. For cancer *vs* non-cancer predictions, the top-performing model included the plasma concentrations of ARG1, CA2, F13A1 and S100A12, achieving a mAUROC of 85.3% (sd = 6.4%). The best model for gastritis *vs* non-gastritis classification combined *H. pylori* status with the plasma concentrations of ARG1,

KRT14, and F13A1, yielding a mAUROC of 82.2% (sd = 9.2%). For preneoplasia *vs* non-preneoplasia prediction, the strongest model was based on the patient age and gender (mAUROC = 75.7%, sd = 8.9%), closely followed by a model that also incorporated the plasma concentration of KRT14 (mAUROC = 74.4%, sd = 10.6%). Thus, predicting preneoplasia *versus* other categories appears more challenging even if the highest mAUROC values exceed 50%, indicating interesting predictive capability.

Full cohort validation. Finally, we assessed the best-performing models on the full cohort to obtain final performance estimates. Using the combinations of biomarkers displaying the best mAUROC values with repeated cross-validations, we assessed their performance to classify patients on the full cohort using linear logistic regression and random forests (Figure 5D). Notably, the AUROC values for the models classifying cancer *vs* non-cancer, and cancer/preneoplasia *vs* healthy/gastritis, were estimated superior to 94.1% and 97.3%, respectively, using random forests (lower bounds of 95% confidence intervals). This demonstrates that the combinations of short-listed biomarkers corresponding to ARG1, CA2, F13A1 and S100A12 for the distinction cancer *vs* non cancer and ARG1, CA2, HPT, MAN2A1, and LBP to distinguish cancer or preneoplasia *vs* healthy or gastritis, enable near-perfect classification of patients for these tasks on the full cohort.

Discussion

GC is a multifactorial disease often diagnosed at an advanced stage, limiting the efficacy of therapeutic interventions and contributing to its poor prognosis. Currently, diagnosis relies on upper endoscopy, an invasive and resource-intensive procedure that requires specialized equipment and highly skilled clinicians to accurately identify gastric preneoplasia, particularly DYS which need a periodical surveillance (35)(36). A current important gap is the lack of non-invasive diagnostic methods for early detection of GC. In this context, the development of blood-based biomarkers represents a promising alternative to overcome the limitations of endoscopic screening and to reduce the global burden of GC.

Changes in the circulating proteome reflect physiological and pathological states, including those associated with inflammation, cellular stress, and malignant transformation. In this study, bottom-up LC-MS/MS-based proteomics of plasma samples from patients with gastric lesions at various stages of the carcinogenesis cascade, followed by ELISA validation, were used to identify and confirm candidate plasma protein biomarkers associated with preneoplastic and cancerous gastric lesions. Our results provide compelling evidence that specific protein concentration profiles in plasma can serve as reliable indicators of disease stage and may be combined to develop non-invasive diagnostic models capable of identifying individuals at high risk of GC, even at asymptomatic stages.

LC-MS/MS analyses highlighted many candidate biomarkers across patient categories (408 differentially abundant proteins in the "relaxed" statistical analysis and 236 in the "strict" analysis). Many of these proteins are involved in pathways known to be dysregulated in cancer, as annotated in KEGG (https://www.genome.jp/kegg/pathway.html) and the Human Protein Atlas (https://www.proteinatlas.org) databases. These include proteins related to signaling and cellular processes (DCD, IL1RAP, KRT14, KRT19, S100A12, SAA1/SAA2, SVEP1, TFRC), innate immunity and immune response (F13A1, IGHG1, LBP), cell motility and membrane

trafficking (CAPZA1, KIF20B), glycan biosynthesis (FUCA2, MAN2A1), digestive and cancer-related systems (CA2, JUP, LEP, PRCP, PRSS3), metabolism, cell energy and growth (ARG1, ATAD3B, ERAP2, GLPD1, HPT, IGFALS), antioxidant system (SOD1). Moreover, most of them are referred as potential prognostic biomarkers in liver (TFRC and GPLD1), renal (IGHG1, PRCP, SAA1/2 and IL1RAP) and digestive (PRSS3) cancers in the Human Protein Atlas database. Importantly, in the GC proteomics database (GCPDB) (37), differential levels of IGFALS, LBP, S100A12, F13A1, SAA1/SAA2, PRCP, IGHG1, have previously been reported in plasma/serum proteome MS analysis on GC patients at various stages. This highlights the relevance of the candidate biomarkers we retrieved from LC-MS/MS experiments.

From the discovery set, fifteen proteins were short-listed for validation by ELISA in a larger patient cohort of 138 subjects among which 106 patients. Four of these proteins (DCD, F13A1, KIF20B and KRT19) did not show significant differences in their plasma concentration between patient categories at the ELISA level, potentially due to limited sample size or differences in detection sensitivity between techniques. In contrast, LEP emerged as a particularly promising biomarker, displaying significant differential abundance in several patient groups, highlighting preneoplasia among others. Due to its diverse functions in the digestive system, including the regulation of immune response, cell proliferation and tissue repair, lipid and glucose metabolism (38), higher levels of LEP have already been associated with gastric inflammation and IM (32), making this protein a biomarker of special interest for gastric preneoplastic and carcinogenesis process. Additional proteins, including ATAD3B, IGFALS, JUP, LBP, and MAN2A1, also displayed significantly higher plasma concentration in the preneoplasia group compared to healthy individuals, suggesting potential utility as early indicators of gastric mucosal transformation. Notably, plasma concentration of several of these proteins was influenced by age and gender, factors that were incorporated into our predictive modeling to enhance diagnostic performance.

When comparing GC patients to healthy individuals, nine proteins (ARG1, ATAD3B, CA2, HPT, IGFALS, JUP, KRT14, LBP and MAN2A1) displayed significantly altered plasma levels, indicating their potential as GC biomarkers. Most of these proteins have already been reported in carcinogenesis process and suggested as diagnosis/prognosis cancer biomarkers. Among these, ARG1 was notably more abundant among men and older GC patients. ARG1 is an enzyme related to metabolic pathways, playing a crucial role in the urea cycle by catalyzing the hydrolysis of arginine to ornithine and urea. It is a potential diagnostic biomarker of cancer progression since it modulates oncogenic signaling pathways such as PI3K/AKT/mTOR, STAT3 and MAPK (39). It has also been mentioned as enriched in hepatocellular carcinoma (40). ATAD3B belongs to the AAA+-ATPase ATAD3 family of mitochondrial membrane-bound proteins, originally reported to be associated with tumorigenesis (41) and suggested as potential prognostic biomarker for hepatocellular carcinoma patients (42). CA2 is a metalloenzyme involved in many physiological process including gastric acid secretion (43). It helps regulate pH, which is vital for cancer cells survival and proliferation notably to adapt to the acidic microenvironment of tumors. Low CA2 expression has been associated with poor prognosis in GC patients (44). HPT showed differential abundance specifically in GC patients *vs* healthy individuals. Serum levels of this hemoglobin binding glycoprotein are associated with inflammatory response and previously proposed as biomarker to predict colorectal cancer and tumor progression (45). In the case of GC, glycomic profiling have reported aberrant HPT glycosylation, suggesting it as a signature molecule for its diagnosis and detection (46). IGFALS is a member of the Leucine-rich repeat sequence (LRR) family and plays a crucial role in the stability and functions of circulating insulin growth factors (IGFs). IGFALS is down-regulated in hepatocellular carcinoma patients and associated with poor patient prognosis (47). JUP is a cell-cell junction protein and a member of catenin families. JUP can be also found in the cytoplasm and nuclei where it acts as a transcription factor, together with ß-catenin. Its loss has been correlated with GC malignancy and

poor prognostics (48). KRT14 is a member of the keratin family of intermediate filament proteins. Changes in KRT14 abundance could be an indicator of the remodeling of the gastric epithelium, promoting tumor growth and invasion (49). LBP has been shown to be intricately linked to GC pathogenesis and proposed as a prognostic biomarker (50) and a diagnostic biomarker for liver metastasis in GC (51). MAN2A1 is an enzyme involved in the processing of N-linked oligosaccharides in the Golgi apparatus. It has been proposed as a potential biomarker significantly related to prognosis and lymph node metastasis in colorectal cancer patients (52). Changes in MAN2A1 abundance might contribute to the metabolic reprogramming and altered glycosylation that are characteristic of cancer cells.

While individual biomarkers showed encouraging trends, the heterogeneity in expression patterns and inter-patient variability prevented accurate classification of patient categories based on the plasma concentration of a single biomarker. Correlation analysis of plasma concentration levels revealed distinct biomarker clusters, suggesting that combining multiple markers could enhance classification accuracy and disease stage differentiation. The benefit of using multiple biomarkers was confirmed by estimating prediction models based on several proteins for different classification tasks using repeated cross-validations. Based on the concentration of the fifteen protein biomarkers, we built prediction models achieving mAUROC values between 75.7% and 88% according to the classification tasks. While patients' age, *H. pylori* status and the plasma concentration levels of ARG1 and KRT14 were key predictors in several models achieving high mAUROC, the optimal biomarker combinations varied depending on the classification task. Among all estimated models, the panel composed of ARG1, CA2, HPT, MAN2A1 and LBP (mAUROC: 83.9%) proved particularly promising as it effectively differentiates patients with cancer or precancerous lesions, from healthy individuals or those with gastritis. An even better diagnostic performance (mAUROC: 85.3%) is achieved with the panel ARG1, CA2, F13A1 and S100A12 for differentiating GC patients from non-cancer individuals.

ARG1 and CA2 are common to both panels suggesting their association with GC as also supported by their mAUROC value among the highest as a single biomarker for the same comparison (75.5% and 78.1, respectively). While S100A12 and F13A1 do not show significant individual concentration differences between GC patients and healthy individuals, they emerge as interesting biomarkers to combine with ARG1 and CA2 for predicting GC. S100A12, a member of calgranulin protein family, is related to innate immunity and has been proposed as a prognostic factor in GC (53). Concerning the fibrin-stabilizing factor F13A1, it has been associated with tumor metastasis in lung cancer (54) and poor prognosis in several cancers (55) (56) (57).

Among the different classification tasks, predicting preneoplasia versus other categories proved to be particularly challenging, with relatively lower mAUROC values than for the other tasks. Although the best models achieved mAUROC values above 74% for this task, indicating some predictive capability, their overall performance suggests that additional biomarkers or alternative modeling approaches may be necessary to enhance discrimination in this category. The heterogeneity of the group defined as preneoplasia including patients with AG either with or not IM and/or DYS can also account of these data. However, this limitation highlights the complexity of the preneoplasia condition, which may exhibit overlapping biomarker profiles with other disease stages, as for advanced GC or gastritis. Indeed, patients with gastritis displayed quantified plasma proteomes highly correlated with those of preneoplasia patients, while biomarkers of preneoplasia can also be biomarkers of cancer. The challenge in distinguishing preneoplasia from other categories suggests the need for further refinement of predictive models, possibly through the inclusion of additional molecular markers (e.g., glycomics, metabolomics) or advanced machine learning techniques.

Conclusion

Our study presents an innovative approach for the early detection of GC and its precursor lesions by leveraging plasma proteomics to identify key protein biomarkers. By incorporating a wide range of patient conditions, including gastritis and preneoplasia, our analysis offers a refined and clinically relevant perspective on early cancer detection, encompassing intermediate disease stages. We developed predictive models that achieved high classification performance using a limited number of biomarker concentration measurements. These models were rigorously validated through repeated cross-validation and further confirmed in the full patient cohort, notably achieving AUROC values exceeding 94% for cancer *vs* non-cancer classification when employing random forest algorithms. These promising results pave the way for future research to further refine these models by incorporating larger and more diverse datasets, as well as external validation in independent, large clinical cohorts. Expanding these efforts will enhance the robustness and generalizability of the models, ultimately facilitating their integration into routine clinical practice for more effective early diagnosis and intervention. This study underscores the transformative potential of plasma proteomics in improving GC detection and management, with the goal of developing a non-invasive and highly predictive blood test for identifying at-risk patients using a limited set of biomarkers.

List of abbreviations

AG : Atrophic Gastritis
AKT : Protein kinase B
AUROC : Area Under the Receiver Operating Characteristics
ATAD3B : ATPase family AAA domain containing 3B
ARG1 : Arginase-1
CA2 : Carbonic anhydrase 2
CA12-5 : Carbohydrate Antigen 12-5
CA19-9 : Carbohydrate Antigen 19-9
CAPZA1: Capping Actin Protein of Muscle Z-Line Subunit alpha
CEA : Carcinoembryonic Antigen
DCD : Dermcidin
DDA : Data Dependent Acquisition
DIA : Data Independent Acquisition
DYS : Dysplasia
ELISA : Enzyme-Linked Immunosorbent Assays
ERAP2: Endoplasmic Reticulum Aminopeptidase 2
F13A1: Coagulation factor XIII A chain
FC : Fold-Change
FDR : False Discovery Rate
FUCA2: Alpha-L-Fucosidase 1
GC : Gastric Cancer
GI : Gastrointestinal
GPLD1 : Glycosylphosphatidylinositol phospholipase D
HPT : Haptoglobin
IGFALS : Insulin-like growth factor-binding protein complex acid labile subunit
IGHG1 : Immunoglobulin heavy constant gamma 1
IL1RAP : Interleukin-1 receptor accessory protein
IM : Intestinal Metaplasia
JUP : Junction Plakoglobin
KIF20B : Kinesin-like protein KIF20B
KRT14 : Keratin, type I cytoskeletal 14
KRT19 : Keratin, type I cytoskeletal 19
LBP : Lipopolysaccharide-binding protein
LC-MS/MS : Liquid Chromatography coupled to tandem Mass Spectrometry
LEP : Leptin
MAN2A1 : Alpha-mannosidase 2A1
MAPK : Mitogen-activated protein kinase
mTOR : Mammalian target of rapamycin
PG : Pepsinogen
PI3K : phosphoinositide 3-kinase
PLS-DA : Partiel Least Square-Discriminated Analysis
PRCP : Lysosomal Pro-X carboxypeptidase
PRSS3 : Serine protease 3
S100A12 : S100-A12 protein
SAA1/2 : Serum amyloid A-1 protein / Serum amyloid A-2 protein
STAT3 : Signal transducer and activator of transcription 3
SVEP1 : Sushi Von Willebrand Factor Type A, EGF And Pentraxin Domain containing 1

TFRC : Transferrin Receptor Protein 1

Declarations

- **Ethics approval and consent to participate**

The study was approved by the Institut Pasteur Clinical Research Coordination Office (Ref 2013-29), the French advisory committee on information processing in Health Research (Ref 14.202) and the French data protection authority: Commission Nationale de l'Informatique et des Libertés (CNIL) (Ref: 2017-204). Patients gave written informed consent for participation before samples were collected.

- **Consent for publication**

Not applicable

- **Availability of data and materials**

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD062318.

The data sets used/or analysed during the current study are also available from the first author and/or the corresponding author on reasonable request.

- **Competing interests**

Aziz Zaanan has had a consulting or advisory role with Amgen, Merck Serono, Sanofi, MSD, Servier, Astellas Pharma, BMS GmbH & Co. KG, Pierre Fabre, AstraZeneca, Daiichi Sankyo, BeiGene, AbbVie, Gilead Sciences and Jazz Pharmaceuticals ; has received support for travel or accomodation from Servier, MSD, Pierre Fabre, Amgen, Merck, Jazz Pharmaceuticals, and Astellas Pharmaceuticals. Julien Taieb has received travel grants and honoraria for speaker or advisory role from: Astellas, Amgen, BMS, Brenus Pharma, Merck Serono, MSD, Natera, Novartis, Ono pharmaceuticals, Oxford bio therapeutics, Pfizer, Pierre Fabre, Proskope, Servier, Sanofi and Takeda.

- **Funding**

This work was supported by grants from ACIP (Pasteur International Concerted Action) grant ACIP2015-10) from Institut Pasteur Paris, the Odyssey Reinsurance Company and Program Flash Maturation (Project INNOV-80-20) from the Institut Pasteur Research Application and Industrial Relations Department to ET.

- **Authors' contributions**

The authors listed above all contributed to this manuscript and approved the final submitted version.

QGG conceptualized, designed and performed the analysis; wrote the draft. VM designed and performed ELISA experiments and analysis. TD conceptualized, designed and performed the proteomic analyses. KN contributed to analyses. AZ, OC, IT, SP, MNU, CJ, JT, DL recruited patients, contributed to plasma collection and cohort characterization. NJ contributed to regulatory procedures and cohort set-up. DL conceptualized and designed the initial project. MM conceptualized and designed the proteomic analyses. ET conceptualized and designed the initial project, contributed to the analyses and wrote the draft. TD, KN, MM, DL contributed to critical revision of the manuscript. MM, ET supervised the study.

- **Acknowledgements**

We thank Pascale Bellamy from the clinical research coordination pole at the Institut Pasteur Paris for her implication in regulatory requirements and cohort setting up. We also would like to thank Sébastien Quesney from the Institut Pasteur Research Application and Industrial Relations Department for his interest in the study and helpful discussions. We are also grateful to the medical staff from the Unit of Hepato-Gastroenterology at the A. Paré hospitals, Boulogne

Billancourt and the Unit of Gastroenterology and Digestive Oncology at the Georges Pompidou European hospitals, Paris for their implication in the study and samples handling. We also thank all the patients that accepted to participate to this study.

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. nov 2018;68:394-424.
2. Mukaisho KI, Nakayama T, Hagiwara T, Hattori T, Sugihara H. Two distinct etiologies of gastric cardia adenocarcinoma: interactions among pH, Helicobacter pylori, and bile acids. Front Microbiol. 2015;6:412.
3. Moss SF. The Clinical Evidence Linking *Helicobacter pylori* to Gastric Cancer. Cell Mol Gastroenterol Hepatol. 2017;3:183-91.
4. Arnold M, Park JY, Camargo MC, Lunet N, Forman D, Soerjomataram I. Is gastric cancer becoming a rare disease? A global assessment of predicted incidence trends to 2035. Gut. 2020;69:823-9.
5. Anderson WF, Rabkin CS, Turner N, Fraumeni JFJ, Rosenberg PS, Camargo MC. The Changing Face of Noncardia Gastric Cancer Incidence Among US Non-Hispanic Whites. J Natl Cancer Inst. 2018;110:608-15.
6. Shen L, Shan YS, Hu HM, Price TJ, Sirohi B, Yeh KH, et al. Management of gastric cancer in Asia: resource-stratified guidelines. Lancet Oncol. 2013;14:e535-547.
7. Correa P. Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res. 1992;52:6735-40.
8. Rugge M, Genta RM, Di Mario F, El-Omar EM, El-Serag HB, Fassan M, et al. Gastric Cancer as Preventable Disease. Clin Gastroenterol Hepatol. déc 2017;15(12):1833-43.
9. Graham DY, Lee YC. Commentary on: the management of patients with gastric intestinal metaplasia. Gut. 2025;74:699-700.
10. Liou JM, Malfertheiner P, Smith SI, El-Omar EM, Wu MS. 40 years after the discovery of *Helicobacter pylori*: towards elimination of *H pylori* for gastric cancer prevention. Lancet. 15 juin 2024;403:2570-2.
11. Ford AC, Yuan Y, Moayyedi P. Long-Term Impact of Helicobacter pylori Eradication Therapy on Gastric Cancer Incidence and Mortality in Healthy Infected Individuals: A Meta-Analysis Beyond 10 Years of Follow-Up. Gastroenterology. 2022;163:754-756.
12. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, Areia M, Leja M, Esposito G, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51:365-88.
13. Dinis-Ribeiro M, Shah S, El-Serag H, Banks M, Uedo N, Tajiri H, et al. The road to a world-unified approach to the management of patients with gastric intestinal metaplasia: a review of current guidelines. Gut. 2024;73:1607-17.
14. Beck M, Bringeland EA, Qvigstad G, Fossmark R. Gastric Cancers Missed at Upper Endoscopy in Central Norway 2007 to 2016-A Population-Based Study. Cancers (Basel). 2021;13:5628.
15. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359:926-30.
16. Desai S, Guddati AK. Carcinoembryonic Antigen, Carbohydrate Antigen 19-9, Cancer Antigen 125, Prostate-Specific Antigen and Other Cancer Markers: A Primer on Commonly Used Cancer Markers. World J Oncol. févr 2023;14:4-14.
17. 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;17:737.
18. Agréus L, Kuipers EJ, Kupcinskas L, Malfertheiner P, Di Mario F, Leja M, et al. Rationale in diagnosis and screening of atrophic gastritis with stomach-specific plasma biomarkers. Scand J Gastroenterol. 2012;47:136-47.
19. Loong TH, Soon NC, Nik Mahmud NRK, Naidu J, Rani RA, Abdul Hamid N, et al. Serum pepsinogen and gastrin-17 as potential biomarkers for pre-malignant lesions in the gastric corpus. Biomed Rep. 2017;7:460-8.
20. Bazin T, Nozeret K, Julié C, Lamarque D, Touati E. Protein Biomarkers of Gastric Preneoplasia and Cancer Lesions in Blood: A Comprehensive Review. Cancers. 2024;16:3019.
21. Shen Q, Polom K, Williams C, de Oliveira FMS, Guergova-Kuras M, Lisacek F, et al. A targeted proteomics approach reveals a serum protein signature as diagnostic biomarker for resectable gastric cancer. EBioMedicine. 2019;44:322-33.
22. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47:D442-50.
23. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006;22:1540-2.
24. Gianetto QG, Wieczorek S, Couté Y, Burger T. A peptide-level multiple imputation strategy accounting for the different natures of missing values in proteomics data [Internet]. 2020 [cited 7 févr 2025]: http://biorxiv.org/lookup/doi/10.1101/2020.05.29.122770
25. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. Schneidman D, éditeur. PLoS Comput Biol. 2017;13:e1005752.
26. Giai Gianetto Q. Statistical Analysis of Post-Translational Modifications Quantified by Label-Free Proteomics Across Multiple Biological Conditions with R: Illustration from SARS-CoV-2 Infected Cells. In: Burger T, éditeur. Statistical Analysis of Proteomic Data [Internet]. New York, NY: Springer US; 2023 [cited 7 févr 2025]. p. 267-302. (Methods in Molecular Biology; vol. 2426). https://link.springer.com/10.1007/978-1-0716-1967-4_12
27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43:e47-e47.
28. Giai Gianetto Q, Combes F, Ramus C, Bruley C, Couté Y, Burger T. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics. 2016;16:29-32.
29. Pounds S, Cheng C. Robust estimation of the false discovery rate. Bioinformatics. 2006;22:1979-87.
30. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44:837.
32. Capelle LG, de Vries AC, Haringsma J, Steyerberg EW, Looman CWN, Nagtzaam NMA, et al. Serum levels of leptin as marker for patients at high risk of gastric cancer. Helicobacter. 2009;14:596-604.
33. Tas F, Karabulut S, Erturk K, Duranyildiz D. Clinical significance of serum leptin level in patients with gastric cancer. Eur Cytokine Netw. 2018;29:52-8.
34. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419.

35. Kim YI, Park JY, Kim BJ, Hwang HW, Hong SA, Kim JG. Risk of metachronous gastric neoplasm occurrence during intermediate-term follow-up period after endoscopic submucosal dissection for gastric dysplasia. Sci Rep. 2020;10:6747.
36. Young E, Philpott H, Singh R. Endoscopic diagnosis and treatment of gastric dysplasia and early cancer: Current evidence and what the future may hold. World J Gastroenterol. 2021;27:5126-51.
37. Ramesh P, Nisar M, Neha null, Ammankallu S, Babu S, Nandakumar R, et al. Delineating protein biomarkers for gastric cancers: A catalogue of mass spectrometry-based markers and assessment of their suitability for targeted proteomics applications. J Proteomics. 3 2024;306:105262.
38. Kim MH, Kim H. Role of Leptin in the Digestive System. Front Pharmacol. 2021;12:660040.
39. Niu F, Yu Y, Li Z, Ren Y, Li Z, Ye Q, et al. Arginase: An emerging and promising therapeutic target for cancer treatment. Biomed Pharmacother. 2022;149:112840.
40. Wang X, Xu Y, Wang R, Dai N, Zhang W, Li F. The significance of arginase-1 expression in the diagnosis of liver cancer: A protocol for a systematic review. Medicine (Baltimore). 2020;99:e19159.
41. Li S, Rousseau D. ATAD3, a vital membrane bound mitochondrial ATPase involved in tumor progression. J Bioenerg Biomembr. 2012;44:189-97.
42. Liu X, Li G, Ai L, Ye Q, Yu T, Yang B. Prognostic value of ATAD3 gene cluster expression in hepatocellular carcinoma. Oncol Lett. 2019;18:1304-10.
43. Kivelä AJ, Kivelä J, Saarnio J, Parkkila S. Carbonic anhydrases in normal gastrointestinal tract and gastrointestinal tumours. World J Gastroenterol. 2005;11:155-63.
44. Hu X, Huang Z, Liao Z, He C, Fang X. Low CA II expression is associated with tumor aggressiveness and poor prognosis in gastric cancer patients. Int J Clin Exp Pathol. 2014;7:6716-24.
45. Ghuman S, Van Hemelrijck M, Garmo H, Holmberg L, Malmström H, Lambe M, et al. Serum inflammatory markers and colorectal cancer risk and survival. Br J Cancer. 2017;116:1358-65.
46. Jeong S, Oh MJ, Kim U, Lee J, Kim JH, An HJ. Glycosylation of serum haptoglobin as a marker of gastric cancer: an overview for clinicians. Expert Rev Proteomics. 2020;17:109-17.
47. Xu L, Xiong L, Chen Y, Chen J, Liu X, Xu Y, et al. IGFALS suppresses hepatocellular carcinoma progression by stabilizing PPAR-γ. Int Immunopharmacol. 2024;143:113414.
48. Chen Y, Yang L, Qin Y, Liu S, Qiao Y, Wan X, et al. Effects of differential distributed-JUP on the malignancy of gastric cancer. J Adv Res. 2021;28:195-208.
49. Zheng Y, Zhao G, Xu B, Liu C, Li C, Zhang X, et al. PADI4 has genetic susceptibility to gastric carcinoma and upregulates CXCR2, KRT14 and TNF-α expression levels. Oncotarget. 2016;7:62159-76.
50. Lv H, Lv M, Guo X, Zhu X, Chao Y, Li D. Lipopolysaccharide-binding protein (LBP): a prognostic biomarker for gastric cancer linked to immune infiltration. BMC Gastroenterol. 2025;25:205.
51. Xie L, Qiu S, Lu C, Gu C, Wang J, Lv J, et al. Gastric cancer-derived LBP promotes liver metastasis by driving intrahepatic fibrotic pre-metastatic niche formation. J Exp Clin Cancer Res. 2023;42:258.
52. Wang Y, Zhao J, Fu G, Sheng C, Zhu J, Zhong T, et al. MAN2A1 predicts prognosis and progression through cancer-related pathways in colorectal cancer. Transl Cancer Res. 2022;11:3686-97.
53. Li D, Zeng Z, Yu T, Qin J, Wu J, Song JC, et al. Expression and clinical implication of S100A12 in gastric carcinoma. Tumour Biol. 2016;37:6551-9.

54. Ercan H, Mauracher LM, Grilz E, Hell L, Hellinger R, Schmid JA, et al. Alterations of the Platelet Proteome in Lung Cancer: Accelerated F13A1 and ER Processing as New Actors in Hypercoagulability. Cancers (Basel). 2021;13:2260.
55. Lee SH, Suh IB, Lee EJ, Hur GY, Lee SY, Lee SY, et al. Relationships of coagulation factor XIII activity with cell-type and stage of non-small cell lung cancer. Yonsei Med J. 2013;54:1394-9.
56. Raval JS, Berg AN, Djokic M, Roth CG, Rollins-Raval MA. Factor XIII Subunit A Immunohistochemical Expression is Associated With Inferior Outcomes in Acute Promyelocytic Leukemia. Appl Immunohistochem Mol Morphol. 2018;26:202-5.
57. Liu Y, Deng H, Song P, Zhang M. Constructing a Glioblastoma Prognostic Model Related to Fatty Acid Metabolism Using Machine Learning and Identifying F13A1 as a Potential Target. Biomedicines. 2025;13:256.

Legends of Figures

Figure 1. Multivariate analysis of the mass spectrometry-based proteomics data. A: Sample plot of the PLS-DA model using 5 components. These components explain 48% of the total inertia. **B:** Pairwise correlations between samples. Shades of blue represent correlation values. Black squares are sub-matrices corresponding to each category of patients. Red squares are the correlation matrices between samples of patients affected by gastritis and preneoplasia (AG/P).

Figure 2. Identification of protein biomarkers from mass spectrometry-based proteomics data. A: Venn diagrams summarizing the 408 biomarkers found in all the comparisons when imposing at least 2 quantified values in a category of patients (« relaxed » analysis): 309 are differential in at least two comparisons. **B:** Venn diagrams summarizing the 236 biomarkers found in all the pairwise comparisons when imposing at least 9 quantified values in a category of patients (« strict » analysis): 149 are in at least two comparisons. **C:** Sample plot of the sparse PLS-DA model with 3 components. These components explain 20% of the total inertia. Each plot represents the samples projected on two axes, each one corresponding to a latent component of the model. **D:** Correlation circle plot of the sparse PLS-DA. The proteins measured by ELISA test are highlighted in black.

Figure 3. ELISA quantified plasma levels of 15 protein candidates from MS analysis to distinguish patient groups. Protein levels were measured by ELISA commercial assays as indicated in the methods section, on all samples of the cohort. Statistical analysis was performed using Mann-Whithey test**.** Only significant differential levels are indicated **(******P<0.0001; ***P<0.001; **P<0.01; *P<0.05)

Figure 4. Correlations between biomarker concentrations measured by ELISA across patient samples. A: Pearson correlation matrix between plasma concentration levels of the biomarkers. **B:** Hierarchical clustering of biomarkers using a correlation-based distance and the Ward method.

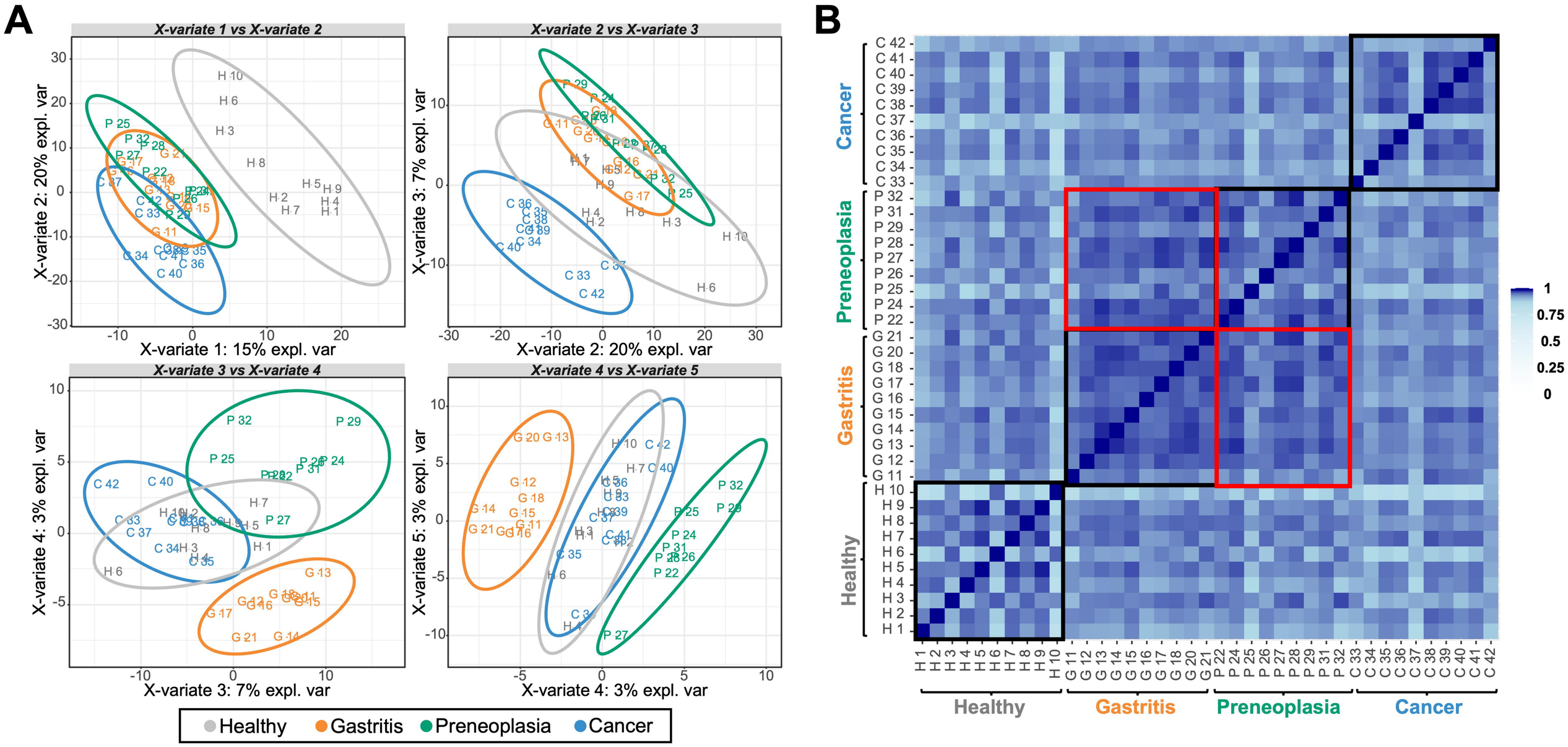
Figure 5. Prediction models with single and multiple biomarkers. A: Schematic representation of the 4-fold cross-validation strategy repeated 10 times, used to evaluate prediction models. **B:** Mean and standard deviation of AUROC for models built using a single biomarker assessed on testing datasets. **C:** Best mean AUROC on testing datasets for models built using multiple biomarkers. **D:** ROC curves assessed for the full cohort from models using biomarkers with the best mean AUROC obtained from repeated cross-validations. 95% CI = 95% confidence interval. RF = random forests, LR = linear logistic regression. AUROC values and ROC curves have been obtained using the pROC R package. Confidence intervals have been obtained with the DeLong methods.

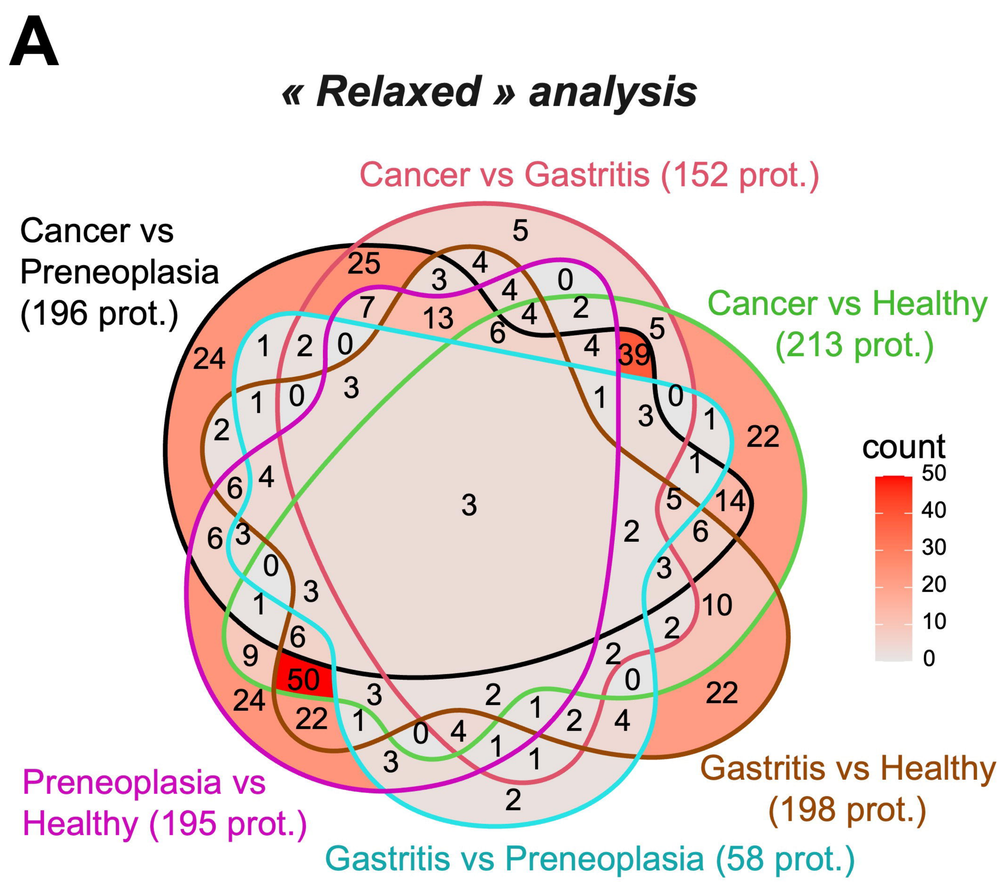
Additional files

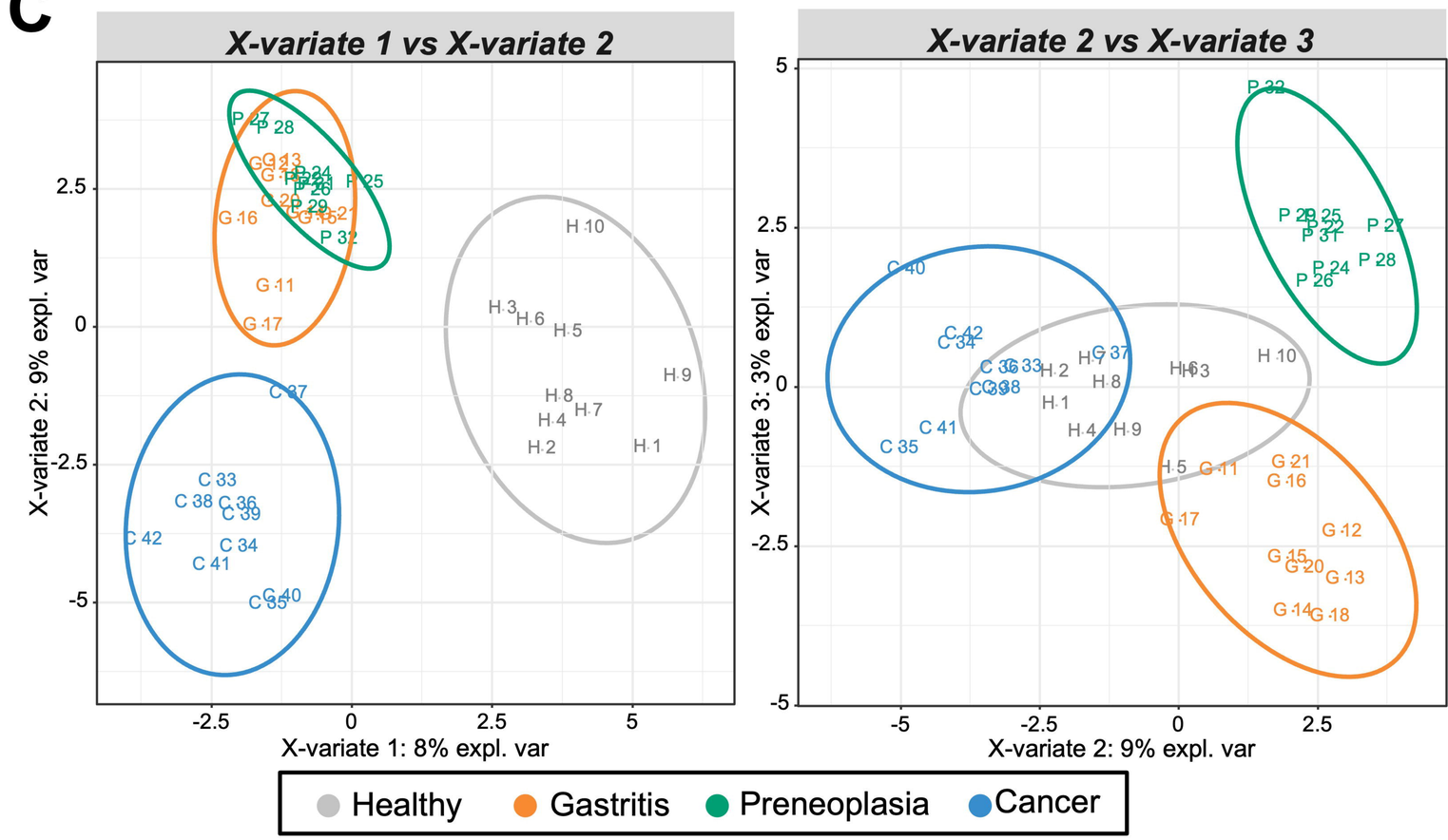
Additional file 1 (xlsx file): Excel file containing the statistical results of the "relaxed" analysis

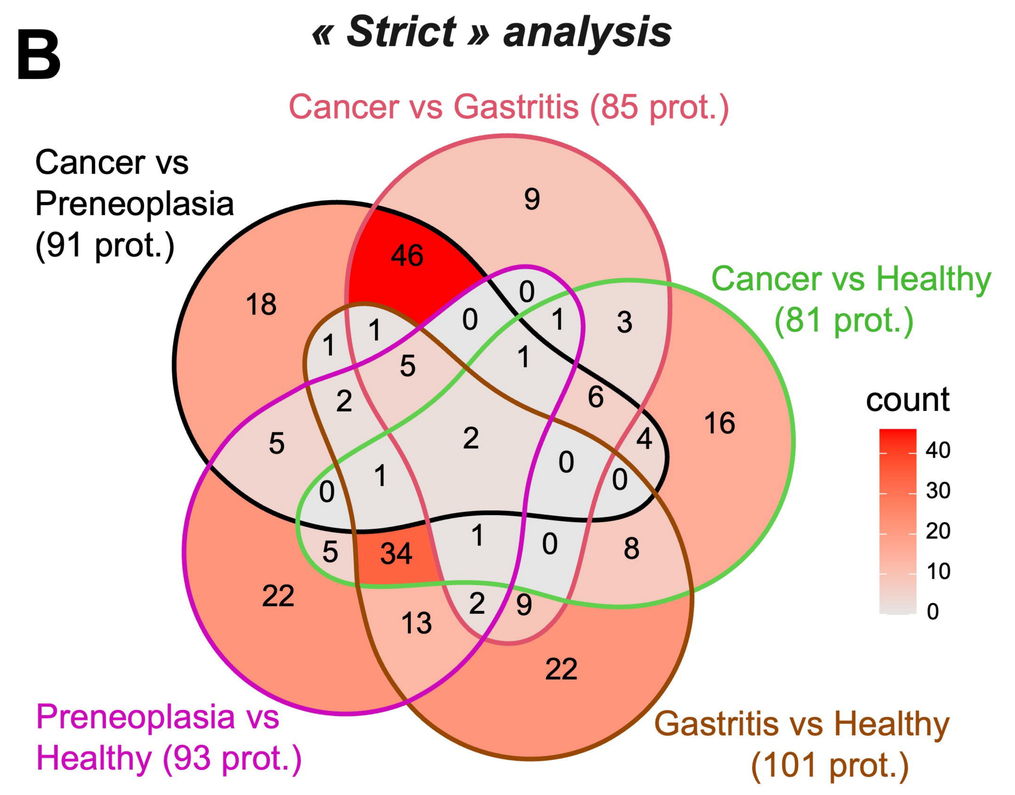
Additional file 2 (xlsx file): Excel file containing the statistical results of the "strict" analysis

Additional file 3 (xlsx file): The means and standard deviations of the Area Under the Receiver Operating Characteristic Curve (AUROC) values obtained through 4-fold cross validation, repeated 10 times from the ELISA data of the 138 patients in the cohort for all tested combinations of biomarkers.

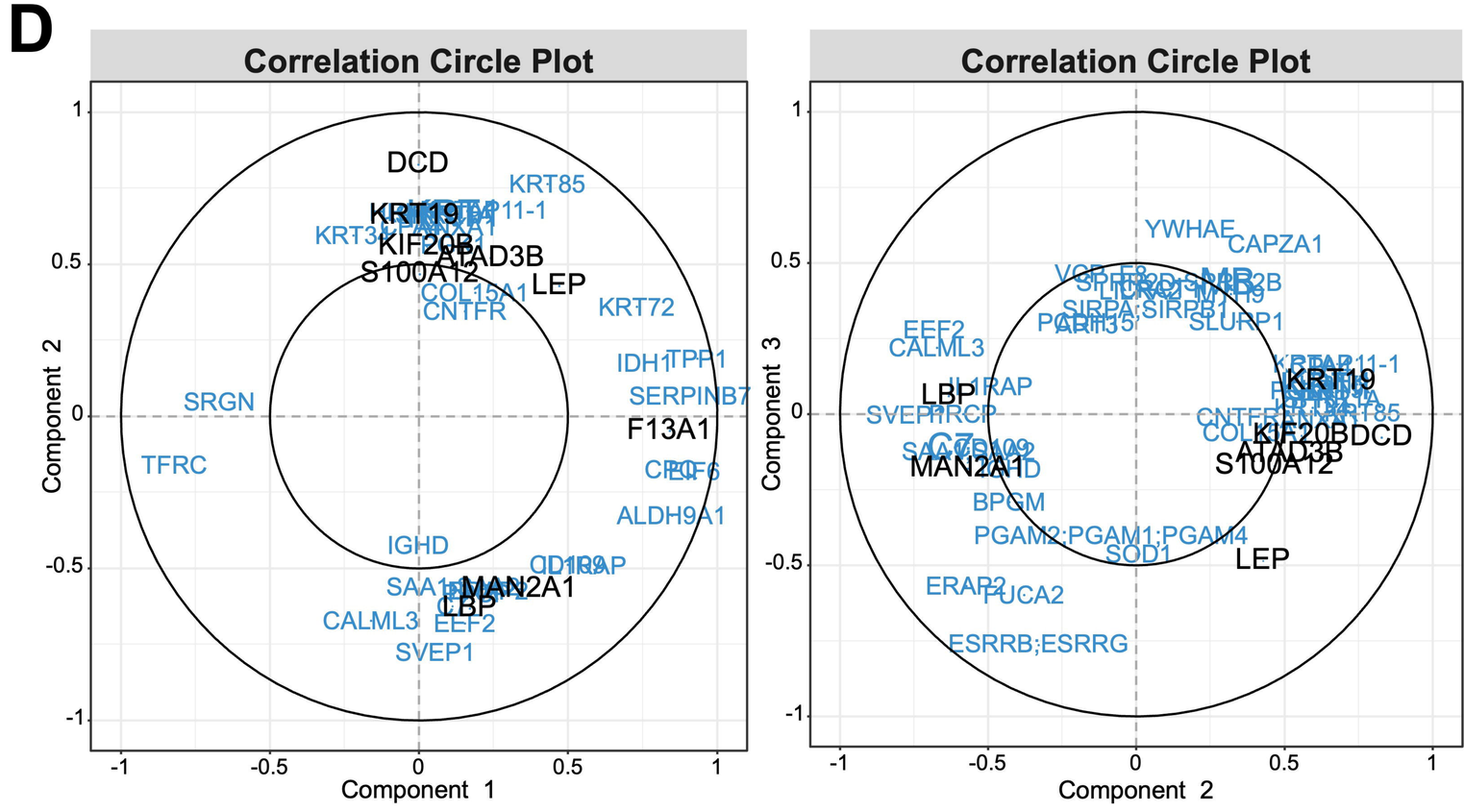

A
X-variate 1 vs X-variate 2
X-variate 1: 15% expl. var
X-variate 2: 20% expl. var
X-variate 2 vs X-variate 3
X-variate 2: 20% expl. var
X-variate 3: 7% expl. var
X-variate 3 vs X-variate 4
X-variate 3: 7% expl. var
X-variate 4: 3% expl. var
X-variate 4 vs X-variate 5
X-variate 4: 3% expl. var
X-variate 5: 3% expl. var
Healthy
Gastritis
Preneoplasia
Cancer
B
Cancer
Preneoplasia
Gastritis
Healthy
1
0.75
0.5
0.25
0

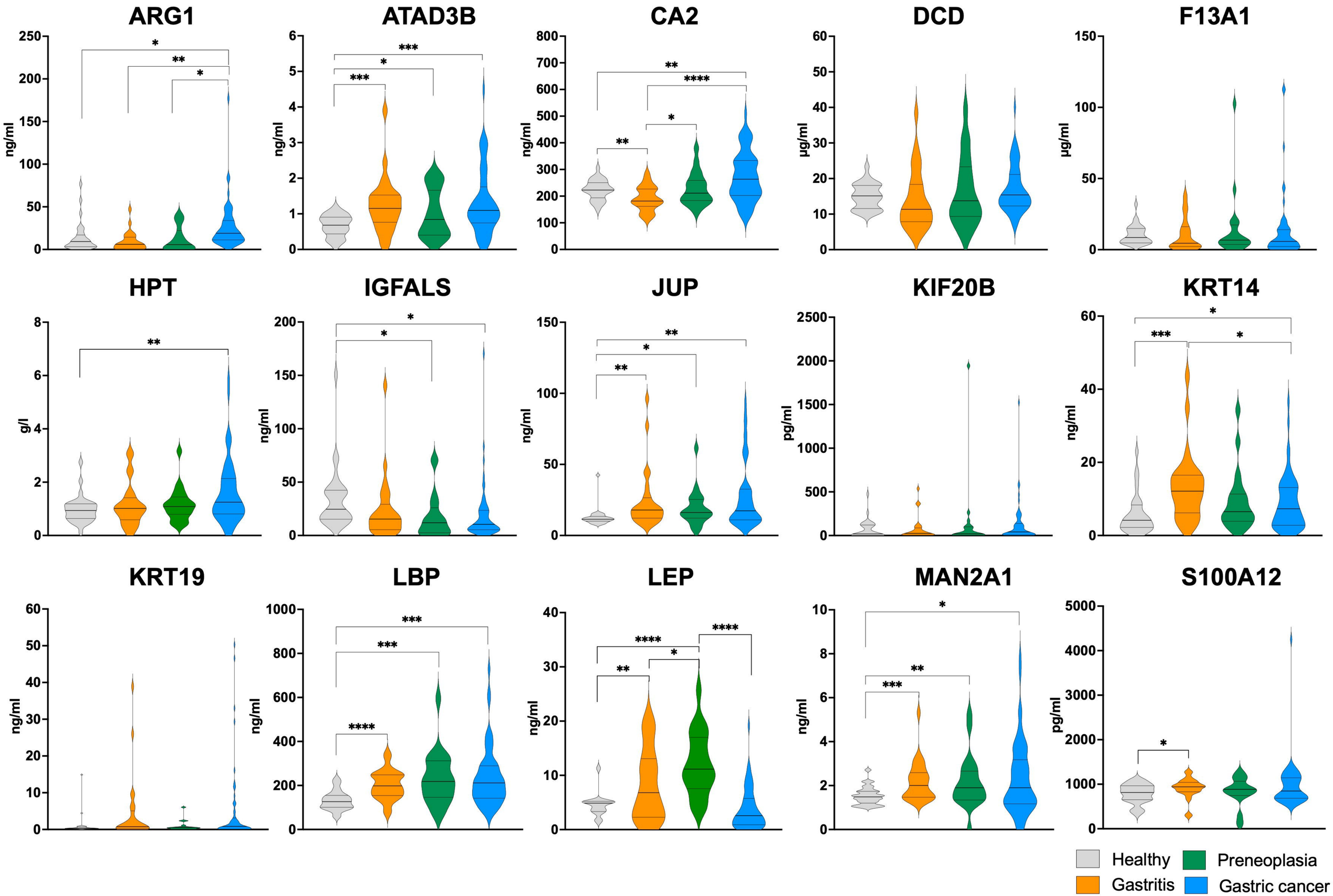






Gastritis vs Preneoplasia: no proteins are differentially abundant

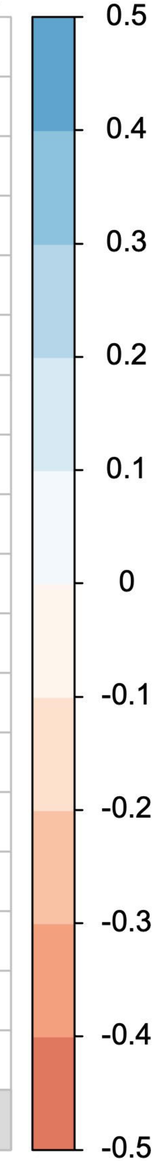



ARG1
ATAD3B
CA2
DCD
F13A1
HPT
IGFALS
JUP
KIF20B
KRT14
KRT19
LBP
LEP
MAN2A1
S100A12
ng/ml
μg/ml
pg/ml
g/l
Healthy
Gastritis
Preneoplasia
Gastric cancer

A

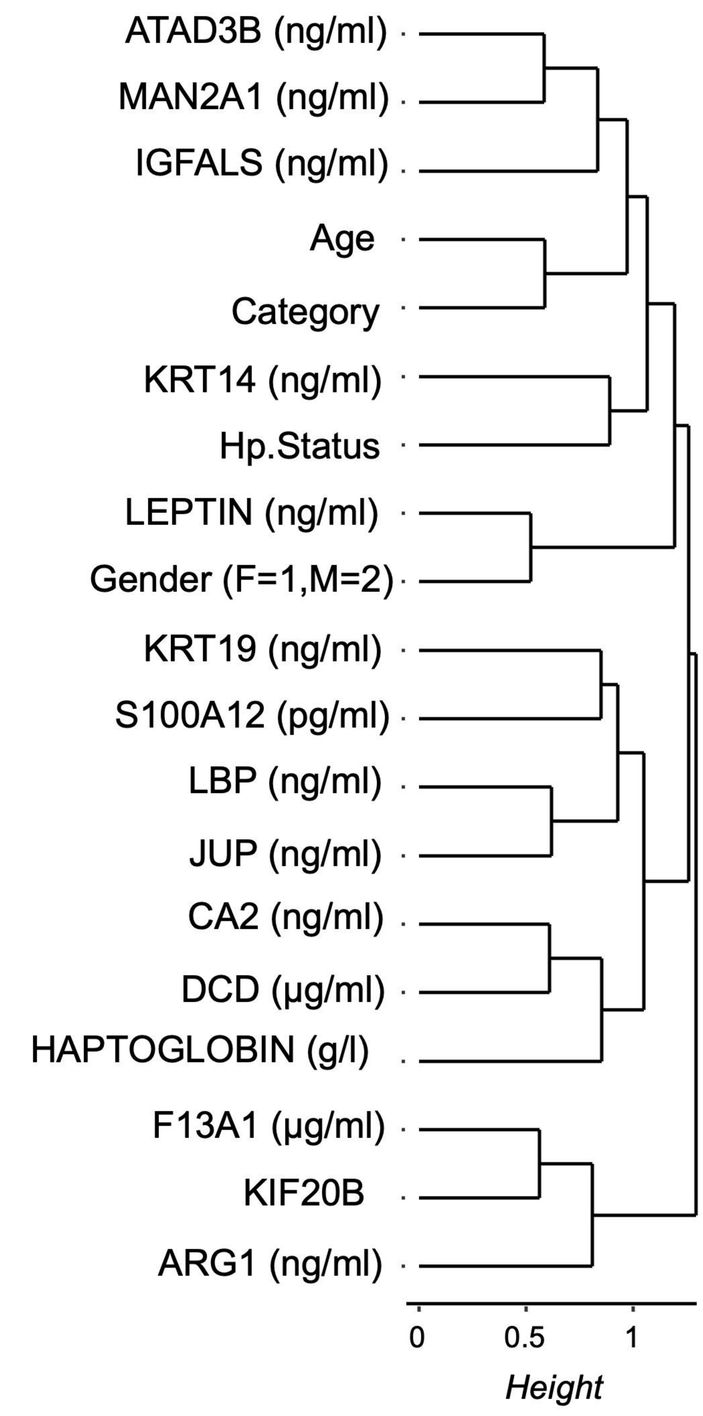
	Category	Age	Gender (F=1,M=2)	Hp.Status	HAPTOGLOBIN (g/l)	LEPTIN (ng/ml)	S100A12 (pg/ml)	KIF20B (pg/ml)	KRT19 (ng/ml)	ARG1 (ng/ml)	MAN2A1 (ng/ml)	IGFALS (ng/ml)	DCD (µg/ml)	ATAD3B (ng/ml)	JUP (ng/ml)	CA2 (ng/ml)	LBP (ng/ml)	F13A1 (µg/ml)	KRT14 (ng/ml)
Category		***	***	*	**					***	***	**		***		**	***		
Age	41		*	*							**								
Gender	31	17				***			***			**						**	
Hp.Status	17	17	5																
HAPTOGLOBIN (g/l)	23	5	8	-1									***	**	*	*	*		
LEPTIN (ng/ml)	-10	7	-48	0	6														
S100A12 (pg/ml)	9	3	4	3	3	-4		*						***		**	**		
KIF20B (pg/ml)	10	2	-11	6	-4	3	18			*	*							***	
KRT19 (ng/ml)	15	9	27	3	13	-13	15	3								***	***		
ARG1 (ng/ml)	34	1	-2	6	4	-14	-4	17	10				*			**		***	
MAN2A1 (ng/ml)	29	20	-4	2	0	3	-13	16	-9	7		***	*	***		***		**	
IGFALS (ng/ml)	-22	-7	-20	-1	1	5	-5	-2	-8	-7	-27		**	*					
DCD (µg/ml)	7	-3	11	7	24	-7	1	-10	4	17	-18	20				***	*		
ATAD3B (ng/ml)	32	9	11	11	22	3	-27	-4	6	0	42	-18	-1			*			**
JUP (ng/ml)	14	13	6	6	16	4	3	6	10	-4	9	1	-2	5		*	***		
CA2 (ng/ml)	24	1	10	-11	16	-9	23	8	28	22	-25	12	39	-18	16		***		
LBP (ng/ml)	28	14	3	4	18	7	23	-4	31	2	-11	5	16	-5	38	29			
F13A1 (µg/ml)	0	14	-19	-1	-3	3	13	44	-8	33	23	0	-9	-2	-4	1	-6		
KRT14 (ng/ml)	6	2	11	11	1	-3	3	-3	14	-8	7	-8	3	22	6	-14	2	-10	

*** $p<0.01$ ** $p<0.05$ * $p<0.1$

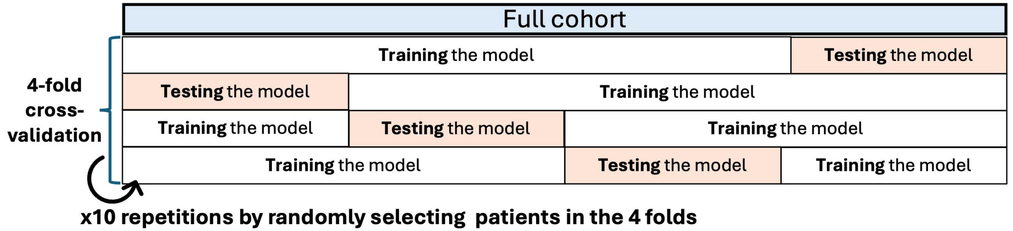


B

Ward's method with an absolute Pearson correlation-based distance



A



B

Mean and standard deviation of AUROC for models built using a single biomarker

	Healthy vs non-healthy		Gastritis vs non-gastritis		Preneoplasia vs non-preneoplasia		Cancer vs non-cancer		Cancer/preneoplasia vs healthy/gastritis		All stages	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Age	**80,1%**	**6,9%**	60,2%	8,1%	**71,8%**	**8,8%**	62,2%	10,0%	**75,1%**	**8,1%**	63,2%	4,8%
DCD (µg/ml)	57,4%	7,0%	60,5%	8,2%	63,7%	8,7%	63,6%	10,2%	55,7%	5,9%	57,1%	5,0%
IGFALS (ng/ml)	63,4%	8,7%	59,6%	8,0%	62,5%	9,3%	58,3%	8,0%	61,2%	7,8%	54,1%	4,9%
LEPTIN (ng/ml)	77,0%	8,5%	65,1%	11,4%	62,4%	8,9%	68,4%	8,8%	63,2%	9,1%	62,7%	6,5%
KRT14 (ng/ml)	63,3%	7,9%	72,3%	8,4%	61,3%	8,0%	58,1%	6,6%	55,8%	5,0%	**63,3%**	**4,4%**
MAN2A1 (ng/ml)	66,0%	8,0%	60,3%	7,8%	60,7%	8,3%	58,8%	8,2%	59,5%	7,4%	57,8%	4,8%
KIF20B (pg/ml)	57,7%	8,2%	57,4%	7,7%	60,4%	7,6%	62,1%	8,8%	57,4%	7,9%	49,9%	5,7%
ARG1 (ng/ml)	60,2%	5,9%	68,0%	10,9%	60,1%	8,6%	75,5%	6,3%	66,1%	7,0%	59,7%	5,9%
F13A1 (µg/ml)	65,6%	7,8%	60,5%	7,1%	60,0%	7,4%	59,1%	6,6%	55,3%	6,0%	50,4%	5,1%
S100A12 (pg/ml)	57,1%	8,4%	59,5%	8,3%	59,9%	8,3%	61,6%	10,9%	58,3%	6,5%	53,8%	5,7%
LBP (ng/ml)	70,7%	6,9%	60,6%	9,3%	59,5%	7,1%	68,4%	9,3%	67,6%	6,2%	61,1%	4,4%
ATAD3B (ng/ml)	61,3%	8,4%	61,0%	7,9%	59,5%	8,0%	59,7%	8,2%	57,3%	8,4%	50,6%	7,4%
JUP (ng/ml)	71,4%	8,1%	58,6%	7,7%	59,3%	6,8%	63,2%	9,9%	65,2%	6,6%	58,8%	5,4%
CA2 (ng/ml)	64,9%	8,6%	63,9%	10,8%	58,8%	7,8%	**78,1%**	**8,7%**	69,4%	7,7%	62,0%	6,1%
HAPTOGLOBIN (g/l)	62,0%	7,9%	61,1%	8,9%	58,5%	6,2%	64,1%	10,0%	64,8%	8,1%	55,6%	4,9%
KRT19 (ng/ml)	63,3%	9,5%	61,0%	9,3%	57,7%	8,5%	60,4%	7,7%	56,0%	7,2%	53,3%	3,8%
HP Status	48,2%	4,9%	67,2%	15,3%	50,9%	3,3%	56,8%	6,1%	54,5%	4,8%	59,6%	3,6%
Gender	57,0%	10,9%	57,7%	7,7%	49,6%	2,4%	66,0%	10,4%	55,4%	9,5%	56,6%	4,0%
max mAUROC	**80,1%**	**6,9%**	**72,3%**	**8,4%**	**71,8%**	**8,8%**	**78,1%**	**8,7%**	**75,1%**	**8,1%**	**63,3%**	**4,4%**
min mAUROC	**48,2%**	**4,9%**	**57,4%**	**7,7%**	**49,6%**	**2,4%**	**56,8%**	**6,1%**	**54,5%**	**4,8%**	**49,9%**	**5,7%**

C

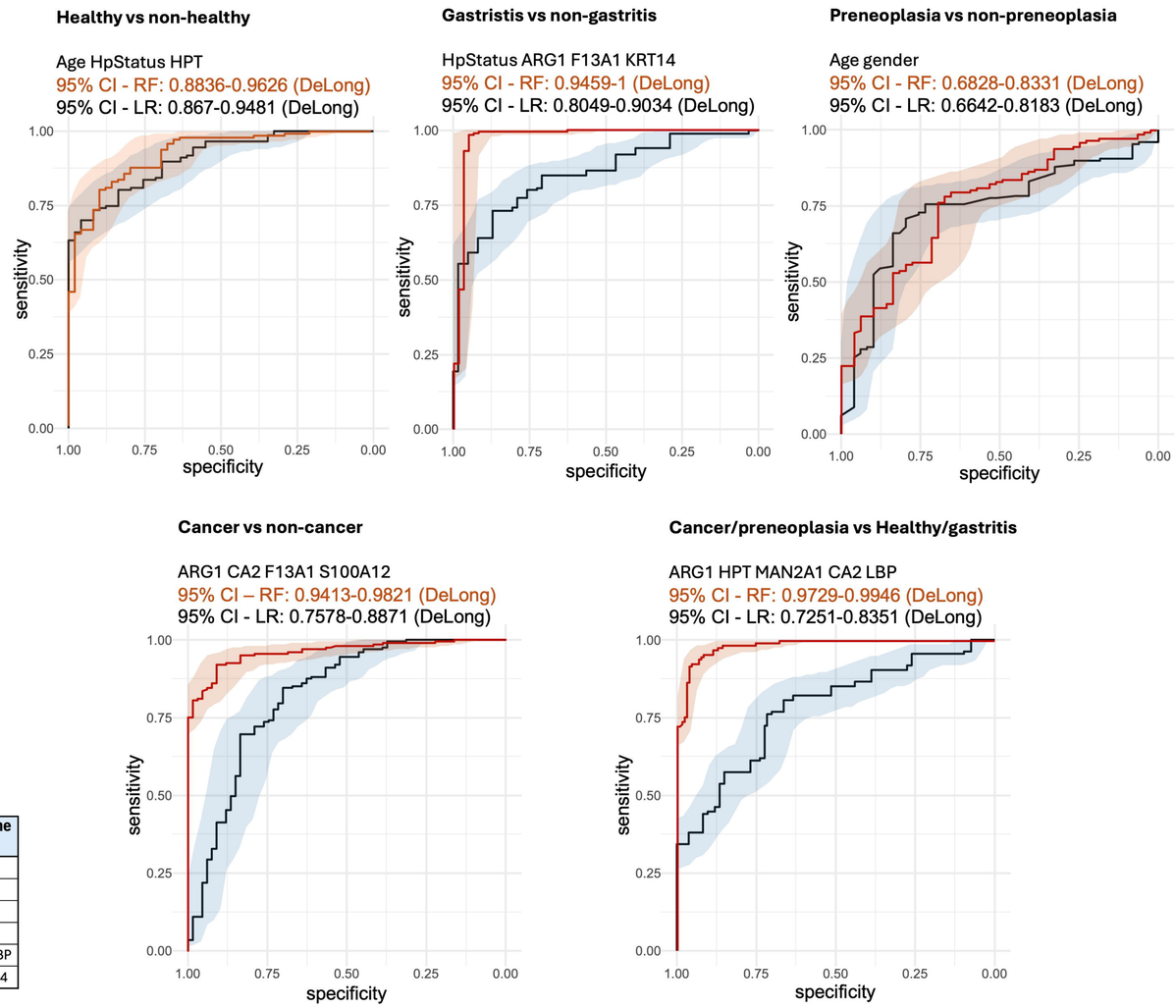
Best mean AUROC on testing datasets for models built using multiple biomarkers

	Nb of biomarkers in the prediction model												Biomarkers in the model with the best mean AUROC
	1		2		3		4		5		6		
	mean	*sd*	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*	
Healthy vs non-healthy	80,1%	6,9%	87,7%	5,0%	**88,0%**	**5,3%**	87,4%	6,9%	86,2%	7,9%	85,2%	8,2%	Age, HP Status, Haptoglobin
Gastritis vs non-gastritis	72,3%	8,4%	81,4%	7,3%	80,6%	7,5%	**82,2%**	**9,2%**	79,6%	9,2%	78,0%	8,0%	HP Status, ARG1, KRT14,F13A1
Preneoplasia vs non-preneoplasia	71,8%	8,8%	**75,7%**	**8,9%**	74,4%	10,6%	73,4%	10,8%	72,7%	7,9%	70,5%	11,3%	Age, Gender
Cancer vs non-cancer	78,1%	8,7%	83,0%	9,4%	85,2%	7,0%	**85,3%**	**6,4%**	85,0%	5,7%	84,9%	7,4%	ARG1, CA2, F13A1, S100A12
Cancer/preneoplasia vs healthy/gastritis	75,1%	8,1%	76,7%	7,7%	78,8%	6,8%	81,2%	6,8%	**83,9%**	**5,1%**	81,4%	6,3%	ARG1, Haptoglobin, CA2, MAN2A1, LBP
All stages	63,3%	4,4%	71,6%	5,4%	74,8%	4,2%	76,2%	6,9%	**77,9%**	**7,7%**	75,6%	6,3%	Age, Gender, HP Status, ARG1, KRT14
Nb of tested combinations*	*18*		*153*		*816*		*3 060*		*8 568*		*18 564*		

*two models tested for each combination: one with random forests, another with multinomial logistic regression

D

ROC curves assessed for the full cohort from models using biomarkers with the best mAUROC



Group	n	Mean age (range)	Sex ratio (M/F)	*H. pylori* positive (%)
Healthy	32 *10*	42 (21-70) *38 (22-69)*	0.4 *0.4*	0 *0*
Gastritis	26 *10*	59 (27-88) *38 (22-69)*	0.7 *0.6*	58% *58%*
Preneoplasia	20 *9*	70 (50-83) *70 (54-83)*	0.4 *0.5*	61% *50%*
Gastric cancer	60 *10*	62 (29-84) *62 (51-74)*	2 *4*	28% *40%*
Total	**138** ***39***			

Table 1: Characteristics of the study population. The gastritis group corresponds to patients with non-atrophic gastritis. The group of preneoplasia is composed of AG (21%), IM (47%) and IM and DYS (32%). GC group includes GC of intestinal type (53%), of diffuse type (37%) and undifferentiated (10%), among which stage II (2%), III (29%) and IV (69%). The exploratory LC-MS/MS analysis was performed on a subgroup of patients (numbers in italic).

Protein name	Uniprot code	Protein description	KEGG pathways	Human Protein Atlas (TCGA)
ARG-1	P05089	Arginase-1	hsa01100 Metabolic pathways hsa00220 Arginine biosynthesis	*Cancer enriched* Liver Hepatocellular Carcinoma. No prognostic found.
ATAD3B	Q5T9A4	ATPase family AAA domain containing 3B	hsa03029 Mitochondrial Biogenesis	*Cancer enhanced* Testicular Germ Cell Tumor. *Potential prognostic* * Liver Hepatocellular Carcinoma, Kidney Renal Clear Cell Carcinoma.
CA2	P00918	Carbonic anhydrase 2	hsa00910 Nitrogen metabolism hsa01100 Metabolic pathways hsa04971 Gastric acid secretion	*Cancer enriched* Kidney Chromophobe Renal Cell Carcinoma. *Potential prognostic** Lung Squamous Cell Carcinoma.
DCD	P81605	Dermcidin	hsa09193 Unclassified: signaling and cellular processes	*Cancer enriched* Breast Invasive Carcinoma. No prognostic found.
F13A1	P00488	Coagulation Factor XIII A Chain	hsa09151 Immune system hsa04610 Complement and coagulation cascades	*Low cancer specificity*. *Potential prognostic** Lung Squamous Cell Carcinoma.
HPT	P00738	Haptoglobin	hsa09181 Protein families: metabolism hsa01002 Peptidases and inhibitors>Serine peptidases hsa04147 Exosomal proteins of other cancer cells	*Cancer enriched* Liver Hepatocellular Carcinoma. *Validated prognostic** Kidney Renal Clear Cell Carcinoma, Stomach Adenocarcinoma Lung Adenocarcinoma. *Potential prognostic*** Hepatocellular Carcinoma.
IGFALS	P35858	Insulin-like growth factor-binding protein complex acid labile subunit	hsa04935 Growth hormone synthesis, secretion and action hsa04147 Exosomal proteins of haemopoietic cells	*Cancer enhanced* Hepatocellular Carcinoma. *Potential prognostic*** Hepatocellular Carcinoma, Lung Adenocarcinoma

JUP	P14923	Junction plakoglobin	hsa04820 Cytoskeleton in muscle cells hsa05200 Pathways in cancer hsa05202 Transcriptional misregulation in cancer hsa05221 Acute myeloid leukemia hsa05226 Gastric cancer hsa05412 Arrhythmogenic right ventricular cardiomyopathy	*Low cancer specificity.* *Potential prognostic*,* Hepatocellular Carcinoma *Potential prognostic*** Kidney Renal Clear Cell Carcinoma.
KIF20B	Q96Q89	Kinesin-like protein KIF20B	hsa04814 Motor proteins hsa04131 Membrane trafficking hsa04812 Cytoskeleton proteins	*Low cancer specificity.* *Potential prognostic** Lung Adenocarcinoma, *Potential prognostic*** Kidney Renal Clear Cell Carcinoma *Validated prognostic** Pancreatic Adenocarcinoma.
KRT14	P02533	Keratin, type I cytoskeletal 14	hsa04915 Estrogen signaling pathway hsa05150 *Staphylococcus aureus* infection hsa04147 Exosomal proteins of epithelial, colorectal cancer, melanoma cells	*Cancer enriched* Head and Neck Squamous Cell Carcinoma. *Potential prognostic*** Breast Invasive Carcinoma *Potential prognostic** Lung Adenocarcinoma.
KRT19	P08727	Keratin, type I cytoskeletal 19	hsa04915 Estrogen signaling pathway hsa05150 *Staphylococcus aureus* infection hsa04147 Exosomal proteins of epithelial, colorectal cancer, melanoma cells	*Low cancer specificity.* *Potential prognostic** Pancreatic Adenocarcinoma. *Validated prognostic** Kidney Renal Clear Cell Carcinoma
LBP	P18428	Lipopolysaccharide-binding protein	hsa04064 NF-kB signaling pathway hsa04620 Toll-like receptor signaling pathway hsa04936 Alcoholic liver disease hsa05152 Tuberculosis hsa05417 Lipid and atherosclerosis hsa04147 Exosomal proteins of other cancer cells	*Cancer enriched* Liver Hepatocellular Carcinoma. *Validated prognostic** Kidney Renal Clear Cell Carcinoma *Potential prognostic** Kidney Renal Papillary Cell Carcinoma.

LEP	P41159	Leptin	hsa04060 Cytokine-cytokine receptor interaction hsa04080 Neuroactive ligand-receptor interaction hsa04081 Hormone signaling hsa04152 AMPK signaling pathway hsa04630 JAK-STAT signaling pathway hsa04920 Adipocytokine signaling pathway hsa04932 Non-alcoholic fatty liver disease	*Cancer enhanced* Breast Invasive Carcinoma. No prognostic found.
MAN2A1	Q16706	Alpha-mannosidase 2	hsa00510 N-Glycan biosynthesis hsa00513 Various types of N-glycan biosynthesis hsa01100 Metabolic pathways	*Low cancer specificity*. *Potential prognostic*** Kidney Renal Clear Cell Carcinoma *Potential prognostic** Thyroid Carcinoma
S100A12	P80511	Protein S100-A12	hsa04990 EF-hand domain-containing proteins > S100 proteins	*Cancer enriched* Head and Neck Squamous Cell Carcinoma. *Potential prognostic** Stomach Adenocarcinoma.

Table 2: List of fifteen proteins from plasma proteomics analysis that were measured by ELISA. Names, KEGG pathways, reactome pathways, and summary informations from the Cancer resource of the Human Protein Atlas (HPA) are reported. The HPA Cancer resource provides information on mRNA expression, including whether a gene is enriched in a particular cancer type (“Cancer enriched”: at least four-fold higher mRNA level in a particular tissue/cell type compared to any other tissues/cell types; “Cancer enhanced”: at least four-fold higher mRNA level in a particular tissue/cell type compared to the average level in all other tissues/cell types; or low cancer specificity), and whether it is (“potential” or “validated”) prognostic for patient survival in each cancer types from The Cancer Genome Atlas (TCGA).*High expression unfavorable ; **High expression favorable