Screening of Potential Biomarkers for Gastric Cancer with Diagnostic Value Using Label-free Global Proteome Analysis

Yongxi Song^{1,#}, Jun Wang^{1,#}, Jingxu Sun¹, Xiaowan Chen¹, Jinxin Shi¹, Zhonghua Wu¹, Dehao Yu¹, Fei Zhang¹, and Zhenning Wang^{1*}

¹ Department of Surgical Oncology and General Surgery, Key Laboratory of Precision Diagnosis and Treatment of Gastrointestinal Tumors, Ministry of Education, The First Affiliated Hospital of China Medical University, 155 North Nanjing Street, Heping District, Shenyang 110001, China

[#] Equal contribution.

* Corresponding author:

E-Mail: josieon826@sina.cn (Wang ZN)

Running title: Song YX et al/ Diagnostic biomarkers in gastric cancer

ORCID

Yongxi Song, 0000-0002-5736-9453

Jun Wang, 0000-0003-0801-9796

Jingxu Sun, 0000-0002-5243-6266

Xiaowan Chen, 0000-0003-1046-0126

Jinxin Shi, 0000-0003-0975-8806

Zhonghua Wu, 0000-0002-5674-2471

Dehao Yu, 0000-0002-9568-880X

Fei Zhang, 0000-0002-9117-7674

Zhenning Wang, 0000-0003-0557-3097

30 Counts of words: 5237
 31 Tables: 1
 32 Figures: 8
 33 Supplementary Figures: 4
 34 Supplementary Tables: 9
 35 Numbers of all the Reference: 45
 36 Numbers of Reference from 2014: 38

37

38 **Abstract**

39 Gastric cancer (GC) is one of the most malignant tumors worldwide. Despite the
 40 recent decrease in mortality rates, the prognosis remains poor. Therefore, it is
 41 necessary to find novel biomarkers with early diagnostic value for GC. In this study,
 42 we present a large-scale proteomic analysis of 30 GC tissues and 30 matched healthy
 43 tissues using label-free global proteome profiling. Our results identified 537
 44 differentially expressed proteins, including 280 upregulated and 257 downregulated
 45 proteins. The ingenuity pathway analysis (IPA) results indicated that the sirtuin
 46 signaling pathway was the most activated pathway in GC tissues whereas oxidative
 47 phosphorylation was the most inhibited. Moreover, the most activated molecular
 48 function was cellular movement, including tissue invasion by tumor cell lines. Based
 49 on IPA results, 15 hub proteins were screened. Using the receiver operating
 50 characteristic curve, most of hub proteins showed a high diagnostic power in
 51 distinguishing between tumors and healthy controls. A four-protein
 52 (ATP5B-ATP5O-NDUFB4-NDUFB8) diagnostic signature was built using a random
 53 forest model. The area under the curve (AUC) of this model was 0.996 and 0.886 for
 54 the training and testing set, respectively, suggesting that the four-protein signature has
 55 a high diagnostic power. This signature was further tested with independent datasets
 56 using plasma enzyme-linked immune sorbent assays, resulted in an AUC of 0.778 for
 57 distinguishing GC tissues from healthy controls, and using immunohistochemical
 58 tissue microarray analysis, resulting in an AUC of 0.805. In conclusion, this study

identifies potential biomarkers and improves our understanding of the pathogenesis, providing novel therapeutic targets for GC.

KEYWORDS

Gastric cancer; Proteomics; Label-free; Bioinformatics; Diagnosis

Introduction

Gastric cancer (GC) is one of the most common malignant, aggressive tumors, causing approximately 723,100 deaths worldwide in 2012 [1], particularly in East Asia [2]. It is a complex disease with histological and etiological heterogeneity [3]. Large genomic variations have been detected in GC patients [4]. Many patients are diagnosed with GC at a late stage due to the asymptomatic nature of the disease [5]. Despite a decrease in mortality rates in recent years, GC prognosis remains poor, with only 28.3% of new cases surviving for more than 5 years [6]. Our understanding of GC pathogenesis and molecular biology has improved, yet it is still necessary to identify novel biomarkers with early diagnostic value, to determine efficient diagnostic methods, and to discover new targets for treating GC.

Cancer development and progression requires molecular alterations at multiple levels including the genome, transcriptome, proteome, and metabolome [7]. In the past decade, numerous studies have examined molecular mechanisms of cancer using genomic and transcriptomic analyses. Protein dynamics are crucial for determining cancer phenotype, and the rapid development of quantitative proteomic approaches for studies on cancer proteomics stimulated investigations characterizing proteogenomic landscapes for many human cancers, including colorectal cancer [8], prostate cancer [9], breast cancer [10], lung adenocarcinoma [11], and ovarian cancer [12]. These efforts promoted the use of mass spectrometry (MS)-based proteogenomics for clinical use [13].

Several recent studies have investigated proteomic aspects of GC. Using the Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) approach combined

with high-resolution MS analysis, a previous study identified 3914 different proteins in six biopsies from different disease stages ranging from chronic gastritis and intestinal metaplasia to gastric adenocarcinoma [14]. Another study examined four GC tissues and four adjacent normal tissues and identified 431 differentially expressed protein (DEPs) using iTRAQ-based quantitative proteomic analysis [15]. This study showed correlations between MTA2 and HDAC1 expression levels and between lymph node metastasis and TNM staging for GC.

One major limitation of current GC proteomics studies is sample size. Small sample sizes introduce bias to findings and result in data inconsistencies. Also, due to individual heterogeneity, paired tumor and healthy control samples from the same patient should ideally be compared when searching for proteomic alterations [16]. Additionally, integrating proteomics and robust bioinformatics methods might help identify potential novel biomarkers with diagnostic power for GC.

In this study, we present a large-scale proteomic analysis of 30 GC tissues and 30 matched healthy tissues using label-free global proteome profiling. This proteomic analysis helped identify 537 DEPs, including 280 upregulated and 257 downregulated proteins. Results of the ingenuity pathway analysis (IPA) indicated that the sirtuin signaling pathway was most activated, whereas oxidative phosphorylation was the most inhibited pathway. Moreover, the most activated molecular function (MF) was cellular movement, including tissue invasion by tumor cell lines. Subsequently, 15 hub proteins were screened based on IPA enrichment results. Using the receiver operating characteristic (ROC) curve, most of these hub proteins had reliable diagnostic potential to distinguish between tumors and healthy controls. After that, a four-protein (ATP5B-ATP5O-NDUFB4-NDUFB8) diagnostic signature was built using a random forest model. The area under curve (AUC) of this model was 0.886 for the testing set, suggesting a high diagnostic potential. Additional independent datasets testing our four-protein signature using plasma enzyme-linked immune sorbent assays (ELISA) yielded an AUC of 0.778 and accuracy of 71.8% to distinguish GC from healthy controls, and visualizing protein expression with tissue microarray analysis yielded an AUC of 0.805. In conclusion, this study identified

118 highly dysregulated proteins and potential biomarkers with potential use in detecting
119 GC. These results further our understanding of GC pathogenesis and identify novel
120 and specific therapeutic targets for this cancer.

121

122 **Results**

123 **Global proteome profiling in the GC cohort**

124 In this study, we conducted an integrated analysis of the global proteome profile of
125 GC (**Figure 1**). Thirty primary tumor tissues and corresponding adjacent healthy
126 tissues were obtained after surgical resection from a total of 30 GC patients at the
127 First Hospital of China Medical University. The neoplastic purity analysis of all 60
128 samples is shown in Supplementary Figure 1A. High-resolution liquid
129 chromatography-tandem MS (LC-MS/MS) was used to identify differences in
130 proteomic profiles between the tumor and healthy samples. LC-MS/MS analyses were
131 performed using MAXQUANT software (v1.4.1.2) [17]. The distribution of peptides,
132 unique peptides, and identified proteins are shown in Supplementary Figure 1B. We
133 used the label-free quantification (LFQ) algorithm embedded in MAXQUANT to
134 quantify protein expression, and peptide-spectrum matching, false discovery rate
135 (FDR), peptide FDR threshold, and protein FDR threshold were all set to 1%. A total
136 of 10,615 proteins were identified in this study, with an average protein coverage rate
137 of 28% (Supplementary Figure 2A and 2B). Of these, expression of 10,576 proteins
138 was quantified, with expression of 2722 proteins quantified across all 30 tissue pairs
139 (Supplementary Figure 2C).

140 We began by calculating the ratio of tumor versus healthy tissues for one paired
141 sample using LFQ values. After that, we generated Spearman's correlation coefficient
142 matrices for all 30 patients using protein ratios (Supplementary Figure 2D). The
143 fraction of total (FOT) value was used to determine the distribution of protein
144 expression profiles across all GC samples (Supplementary Figure 2E). Our results
145 indicated consistent proteome identification and quantification throughout our study.

We next determined the coefficients of variation (CVs) and interquartile range (IQR) of the proteins. The overall CV decreased significantly when the FOT value was higher than 10^{-5} (Supplementary Figure 3A). Additionally, the increased performance of the IQR was discontinued when the FOT value was higher than 10^{-5} (Supplementary Figure 3B), suggesting that the most suitable value for accurate quantification was when the FOT value was 10^{-5} . This result is consistent with the cut-off value from a previous GC proteomics study [16]. We also calculated the distribution of the quantitative samples and found a median CV of $\log_{10}(\text{FOT})+7$. The median CV was significantly decreased when the number of quantitative samples was over 20 (Supplementary Figure 3C). Therefore, we filtered 3940 proteins using the above cut-off value and the number of quantitative samples (Supplementary Table 2). For these proteins, we used principal component analysis (PCA) to analyze expression in tumor tissues and corresponding matched healthy tissues (Supplementary Figure 3D). The proportion of variance of PC1, PC2, and PC3 was 19.06%, 6.91%, and 5.49%, respectively. These PCA results indicate a clear distinction between the proteomes of tumor and healthy tissues.

162

163 Identification of DEPs in GC

We next assessed significant quantitative differences between tumor tissues and matched healthy tissues. DEPs were screened using a filter criterion of $|\log_2\text{-fold change (FC)}| > 1$ and $\text{FDR} < 0.05$. This identified 537 DEPs, including 280 upregulated and 257 downregulated proteins (**Figure 2A** and Supplementary Table 3). A volcano plot showing statistically significant DEPs between tumor and healthy tissues was constructed (**Figure 2B**). Expression levels of the top ten significant DEPs - DMBT1, SPB5, CPSM, KI67, CEAM5, ATP4A, ATP4B, CLIC6, KCRB, and LIPG - are shown in **Figure 2C** and **Table 1**. Subcellular localization analysis indicated that a large number of DEPs were annotated as mitochondrial, suggesting that the GC proteome is involved in tumor energy metabolism (**Figure 3A**). The gene ontology (GO) term enrichment analysis (Supplementary Table 4) revealed that upregulated DEPs in GC were significantly enriched in activities of the nucleolus, DNA, and RNA.

176 Downregulated DEPs were mostly enriched in mitochondrial processes, cellular
177 respiration, and oxidative phosphorylation (**Figure 3B**).

178

179 **Proteomic pathways and potential hub proteins in GC**

180 DEPs were further categorized using the IPA database to identify proteins with
181 potential significant diagnostic values for GC. The pathway enrichment analysis
182 indicated that DEPs were significantly enriched in oxidative phosphorylation,
183 mitochondrial dysfunction, the sirtuin signaling pathway, and the tricarboxylic acid
184 (TCA) cycle (Supplementary Table 5). Of the 25 significant signaling pathways, five
185 pathways (sirtuin signaling, interferon signaling, inflammasome, IL-8 signaling, and
186 neuroinflammation signaling) with z-scores > 0 were predicted to be activated in GC
187 whereas 20 pathways (z-score < 0) were predicted to be inhibited (**Figure 4A**). The
188 most inhibited pathways were oxidative phosphorylation and the TCA cycle. Of
189 particular interest, oxidative phosphorylation is potentially inhibited and the sirtuin
190 signaling pathway activated in GC (**Figure 4B**). All proteins identified to be involved
191 in oxidative phosphorylation were downregulated.

192 We performed disease and functional analyses for the above DEPs in IPA (**Figure**
193 **4C**), revealing that the most activated function was cellular movement, including
194 tissue invasion by various tumor cell lines. Tumor cell line proliferation, adhesion,
195 and inflammatory responses were also found to be activated in tumors. The most
196 inhibited function was lipid export. The network associated with regulation of tissue
197 invasion by tumor cell lines and lipid export was connected by five proteins: ACAT1,
198 CAV1, CTSS, S100A12, and S100A9 (**Figure 5A**).

199 Our biomarker analysis identified 72 potential biomarkers associated with cancer
200 and gastrointestinal diseases, and the network analysis implicated 25 networks.
201 Significant regulatory networks with scores > 30 were found associated with cell
202 signaling, post-translational modification, protein synthesis, protein trafficking,
203 energy production, nucleic acid metabolism, small molecule biochemistry, and
204 cell-to-cell signaling and interaction (**Figure 5B**).

205

206 Screening potential diagnostic markers in GC

207 We next screened fifteen hub proteins from the enriched canonical pathways,
208 biomarker analyses, and the top ten protein-protein interaction networks
209 (Supplementary Table 6). The diagnostic performance of each protein was assessed
210 using the ROC curve. Most proteins (13 out of 15) showed a high AUC (> 0.80)
211 between GC and healthy tissues, suggesting that these hub proteins might have
212 discriminating potential as GC diagnostic markers (**Figure 6A**). Among these,
213 NDUB8 (O95169) and CX7A2 (P14406) had AUC of 0.98 and 0.978, respectively.
214 The 95 % confidence interval (CI) of the AUC and p-values compared with the
215 reference line are shown in Supplementary Table 7.

216 To test our newly identified proteins, we used an independently published cohort
217 of 60 GC samples and matched healthy samples [16]. A total of 561 DEPs including
218 345 upregulated and 216 downregulated proteins were selected using the same filter
219 criteria (Supplementary Table 8). When compared with the results of our initial cohort
220 (**Figure 6B**), 264 DEPs were shared between the two datasets, and 8 of the 15 hub
221 proteins discussed above were differentially expressed in the testing phase. As shown
222 in **Figure 6C**, ATP5O (P48047, AUC = 0.89) and NDUB4 (O95168, AUC = 0.873)
223 were the two proteins with the highest predictive power. Proteins with AUC > 0.70
224 were considered potential independent diagnostic biomarkers (**Figure 6D**). Only one
225 protein (P09669) was excluded. Expression profiles of significant hub proteins found
226 in the testing phase are shown in **Figure 6E**.

227

228 Establishment and validation of a four-protein signature

229 Although the above results indicate that single proteins may hold significant
230 discriminating potential, we investigated the possibility of building a multi-protein
231 signature with increased diagnostic potential, sensitivity, and specificity. We used a
232 random forest model (**Figure 7A**) including the performance of each protein. The
233 best-performing proteins were found to be NDUB4 (O95168), ATPB (P06576), ATPO
234 (P48047), and NDUB8 (O95169) (**Figure 7B**). After increment feature selection,
235 AUC, sensitivity, specificity, and accuracy were found to be constant and suitable

when the four-protein signature was built (**Figure 7C**). The AUC of this protein signature in the training set was 0.996 (**Figure 7D**, red line) and the accuracy was 98.3 %. Thus, we built a four-protein signature (ATP5B-ATP5O-NDUFB4-NDUFB8) with high diagnostic potential for GC.

To test the stability and diagnostic power of the four-protein signature, we analyzed GC and healthy tissues using LC-MS/MS, a testing cohort consisting of proteomic dataset from 60 GC patients [16] using label-free analysis. The AUC of the four-protein signature was 0.886 and its accuracy was 80.0 %, suggesting that this protein panel had a high diagnostic value for GC (**Figure 7D**, blue line).

The above results were obtained from tumor and healthy tissues collected from patients and analyzed using LC-MS/MS. To further extend our study for clinical use and noninvasive detection, we used blood plasma samples from GC patients and healthy individuals to evaluate the potential diagnostic value of the protein signature for detecting GC using blood. We performed ELISA on plasma samples from 37 GC patients and 34 healthy controls for the validation phase. Our results showed that the protein signature had an AUC of 0.778 and accuracy of 71.8 % to distinguish GC tissues from healthy controls (**Figure 7E**).

Finally, we tested 251 pairs of GC tissues using tissue microarray. As shown in **Figure 8A-D**, we first determined differential expression patterns for the four proteins by immunohistochemical staining on tumor and normal tissues (p -values < 0.05). As this dataset included an adequate number of samples, we could divide the samples into internal training (80% samples) and testing (20% samples) sets. The AUC for the internal training set was 0.805, with an AUC of 0.727 for the internal testing set (**Figure 8E**). The p -values compared with reference line for all datasets are shown in Supplementary Table 9.

At the same time, we tested the AUC values of each individual protein from the four-protein combination in above datasets. We found very small AUC values (Supplementary Figure 4). In sum, our results suggest that this novel four-protein signature (ATP5B-ATP5O-NDUFB4-NDUFB8) has a high diagnostic power for GC.

266 **Discussion**

267 In this study, we present a large-scale proteomic analysis of GC using label-free
268 global proteome profiling based on 30 GC tissues paired with 30 healthy tissues. A
269 four-protein (ATP5B-ATP5O-NDUFB4-NDUFB8) diagnostic signature was built
270 using a random forest model. Our findings might help to understand the pathogenesis
271 of GC and provide novel and specific therapeutic targets for this disease.

272 **Biomarkers and proteomics in cancer**

273 Biomarkers may play important roles in the diagnosis and treatment of cancers by
274 enabling early detection and diagnosis [18]. Furthermore, the rapid development of
275 quantitative proteomic approaches has allowed researchers to analyze biomarkers for
276 human tumors. For instance, proteomics were used for biomarker discovery for
277 colorectal cancer, including the role of protein phosphorylation and cancer stem cells
278 [19]. Clinical proteomics has also proven to be a promising tool for improving
279 personalized medicine for colorectal cancer using blood, stool, and biopsy samples
280 [20]. Moreover, MS-based proteomics have been used for drug discovery and
281 development [21].

282 Other groups have also performed proteomic studies for GC. Huang et al.
283 performed a quantitative proteomic study using ten GC serum samples and healthy
284 controls using tandem mass tags [22], and identified 594 serum DEPs with a cut-off
285 value of 1.2-FC. The DEPs 1C12, PIGR, S10A8, AOC3, FHL1, GGCT, NCAM1, and
286 SYNEM were also identified in our study. Liu et al. performed label-free LC-MS/MS
287 using GC tissues and healthy tissues from six patients [23] and found 87 DEPs. Of
288 these 87 proteins, ATPB, ATP4B, NDUB9, and NDUAD were also identified in our
289 study. However, we used a more stringent screening criteria with a cut-off value of
290 2-fold change and an FDR < 0.05. The analysis of our testing set (60 pairs) revealed
291 that 264 DEPs were shared between the two datasets. The DEPs identified by us are
292 different from those of previous studies, likely due to the different platforms,
293 quantitative methods, and screening criteria used. Therefore, more accurate analytical
294 methods and a larger number of samples are necessary to confirm our findings.

Diagnostic test performance is often assessed by measuring ROC, AUC, sensitivity, and specificity. For an example, a label-free quantitative proteomic study was performed to diagnose periodontal diseases using saliva [24], and the authors found that 12 proteins presented the highest AUC (AUC 0.83-0.91) between healthy and diseased tissues. Yang et al. used targeted proteomics coupled with immunoaffinity enrichment to investigate epithelial ovarian cancer samples and found that the combined AUC for serum carbohydrate antigen 125 (CA125) and heat shock protein 27 was 0.88, which was significantly higher than that of CA125 alone [25]. Jiang et al. performed iTRAQ labeling and LC-ESI-MS/MS to evaluate a discovery group of four GC samples and four adjacent healthy tissues [26], and found an AUC of 0.734 for GLS1 and GGCT co-expression, suggesting that the level of co-expression had high clinical value as a diagnostic biomarker for early GC.

IPA pathway analysis

Our results highlighted five pathways (sirtuin signaling, interferon signaling, inflammasome, IL -8 signaling, and neuroinflammation signaling) with z-scores > 0 predicted to be activated in GC. Among these, the sirtuin signaling pathway was the most activated. Sirtuins are members of the class III histone deacetylase family of enzymes [27], and mammalian sirtuins are classified into seven groups (SIRT1-7) [28]. Sirtuin signaling pathways regulate stem cell functions crucial for normal embryonic development and adult tissue homeostasis [29]. In cancer, sirtuins are implicated in producing cancer cells capable of self-renewal and differentiation, resulting in tumor growth. Several studies also confirmed the role of sirtuins in stimulating epithelial-mesenchymal transition [30, 31]. The interferon signaling pathway and other innate immune signaling pathways in tumor cells were shown to be determinants of treatment response and resistance [32]. These pathways may serve as an alternative immunotherapeutic strategy in GC [33, 34], and our results suggest that these activated pathways may play important roles in the pathogenesis of GC.

TCA cycle and oxidative phosphorylation

The most inhibited pathways in GC were oxidative phosphorylation and the TCA cycle. This is consistent with another report on GC [14]. Recent studies have shown

that cancer cells exhibit significant metabolic changes in mitochondrial dynamics and function [35]. Tumor growth is regulated by the TCA cycle and oxidative phosphorylation in mitochondria. According to the Warburg Effect [36], cancer cells are likely to increase energy metabolism via aerobic glycolysis rather than oxidative phosphorylation. Therefore, the regulation of energy metabolism via oxidative phosphorylation and the TCA cycle tends to be inhibited in GC, with aerobic glycolysis stimulated.

The four proteins of the GC signature

Our study identifies a four-protein signature with diagnostic value for GC (ATP5B-ATP5O-NDUFB4-NDUFB8). Of these, ATP5B has not been reported upon in previous studies on GC, however, it was found to be downregulated in clear cell renal cell carcinoma [37]. In glioblastoma, ATP5B mRNA levels were significantly higher in tumor cells compared with healthy brain blood vessels, and microvascular proliferation was significantly higher [38]. ATP5O was reported as one member of 8 mitochondrial proteins (NDUFS5, VDAC3, ATP5O, IMMT, MRPL28, COX5B, MRPL52, PRKDC) which generated a compact gastric mitochondrial gene signature for predicting tumor progression and overall survival of GC patients [39]. ATP5O Gene expression was also downregulated in clear cell renal cell carcinoma [37]. NDUFB8 was found to be hypermethylated in glioblastoma [40]. However, NDUFB4 has been rarely reported upon in human cancers.

ATP5B and ATP5O are ATP synthases, and NDUFB4 and NDUFB8 are two members of the NADH dehydrogenase (ubiquinone) 1 beta sub-complex. Mitochondrial dysfunction is common in cancer, and mitochondrial electron transport chains are often affected in carcinogenesis [41, 42]. Mitochondrial dysfunction is involved in cancer cell metabolism, apoptosis and autophagy. Using antibiotics as anticancer drugs has been considered as potential anticancer therapy [43]. Therefore, targeting mitochondrial alterations might be a promising strategy for the development of tools for GC diagnosis, prognosis and treatment.

The performance of protein signature in plasma

354 The performance of our protein signature was found to be better when measured in
355 tissues (AUC = 0.996 and 0.886) than in plasma (AUC = 0.778). A major reason for
356 this might be due to the stability of different experimental tests including LC/MS-MS
357 and ELISA. A second reason might result from the complexity of the multi-step
358 process for analyzing blood plasma samples from tumor and healthy controls.
359 However, the blood plasma test is still of value. Moreover, as described above, our
360 protein signature closely associated with energy metabolism. Therefore, this
361 diagnostic protein signature may be applicable to other cancers in addition to GC.

362 **The differences between our and Qin's study**

363 Recently, Qin's group presented a dataset providing information on proteomics of
364 gene products and mutations in cancer driver genes from 84 diffuse-type GC patients
365 [16]. They divided the patients into three molecular subtypes that provided a wealth of
366 information on diffuse GC signaling pathways and demonstrated the benefits of
367 proteomic analysis in cancer molecular subtypes. In contrast to with Qin's study, our
368 project focuses on mass spectrometry-based proteomics and bioinformatic algorithms
369 to screen biomarkers and models with diagnostic value in GC. In our study, we
370 identified hub proteins with high diagnostic power in distinguishing tumors and
371 normal controls. We further built a four-protein (ATP5B-ATP5O-NDUFB4-NDUFB8)
372 diagnostic signature using a random forest model then verified it GC tissues as well as
373 with two independent datasets of plasma and tissue microarrays. Our study identified
374 potential biomarkers and may help increase our understanding of GC pathogenesis
375 and provide novel and specific diagnostic targets for this cancer.

376 As shown in our multiple rounds of testing, our narrowed-down four-protein
377 signature had high diagnostic power between tumor tissues and healthy controls,
378 suggesting its potential use as a novel clinical biomarker for GC. However, further
379 large-scale validation studies are necessary to confirm this finding.

380

381 **Conclusion**

382 In summary, this study increased our understanding of GC pathogenesis and identified
383 potential biomarkers to provide novel and specific therapeutic targets for this cancer.

384

385 **Materials and Methods**

386 **Selection of specimens and clinical information**

387 Thirty tumor specimens and 30 matched healthy tissues were obtained from 30 GC
388 patients following surgical resection at the First Hospital of China Medical University
389 (Shenyang, China). Tissue sections were stained with hematoxylin and eosin to
390 evaluate tumor purity by a certified pathologist in the hospital. Tumor samples
391 consisting of at least 60% cancer cells were retained for further analysis
392 (Supplementary Figure 1A and Supplementary Table 1). Each sample was collected
393 within 30 min after surgical resection, cleaned, transferred to sterile freezing tubes,
394 and cryopreserved in liquid nitrogen until further use. All samples were staged
395 according to the seventh edition of The American Joint Committee on Cancer staging
396 system.

397 In the testing phase, the diagnostic value of candidate proteins was estimated
398 using the plasma of 71 individuals (37 GC patients and 34 healthy controls). In brief,
399 2 mL of whole peripheral venous blood were collected from each sample and
400 transferred to a purple-top EDTA tube. Plasma samples were separated following a
401 two-step centrifugation protocol (3000 rpm for 5 min at 4 °C, and 12,000 rpm for 15
402 min at 4 °C) within 4 h after collection, then samples were transferred to
403 RNase/DNase-free tubes and stored at -80 °C until further use.

404 This study was conducted according to the principles established by the
405 Declaration of Helsinki and was approved by the Human Research Ethics Committee
406 of China Medical University (Shenyang, China). A written informed consent was
407 obtained from all patients.

408

409 **Protein extraction and tryptic digestion**

GC tissues were ground into powder in liquid nitrogen and suspended in an ice-cold lysis buffer [8 M urea (V900119-500G, VETEC, China), 5 mM dithiothreitol (DTT, D9163-5G, SIGMA, Canada), 1 % (v/v) protease inhibitor cocktail (539134, Merck, Darmstadt, Germany), 3 μ M trichostatin A (TSA, V900931-5MG, VETEC, China), 50 mM nicotinamide (NAM, N0636-500G, SIGMA, USA)] with occasional sonication. Cell lysates were centrifuged at 12,000 $\times$ g at 4 °C for 10 min, and the resulting supernatants were collected. Total protein concentration was determined from lysates using the 2-D Quant kit (80-6483-56, GE Healthcare, USA) according to the manufacturer's instructions. Proteins were then precipitated with 15 % trichloroacetic acid (TCA, T4885-1KG, SIGMA, China) for 4 h at 4 °C, and the resulting precipitate was washed three times with cold (−20 °C) acetone. Dried protein pellets were resuspended in 100 mM tetraethylammonium bromide (TEAB, T7408-500ML, SIGMA, Switzerland) and digested with trypsin (V5111, Promega, USA) at an enzyme-to-substrate ratio of 1:50 for 12 h at 37 °C. Peptides were reduced with DTT and alkylated with iodoacetamide (V900335-5G, VETEC, China) in the dark. Complete digestion was ensured by performing a second digestion with trypsin at an enzyme-to-substrate ratio of 1:100 for 4 h at 37 °C.

427

428 **Peptide fractionation using high performance liquid chromatography (HPLC)**

429 Samples were fractionated using high pH reverse-phase HPLC using an Agilent
430 300Extend C18 column (particle size, 5 μ m; ID, 4.6 mm; length, 250 mm. Agilent
431 Technologies, USA). Briefly, peptides were first separated into 80 fractions using a
432 gradient of 2 % to 60 % acetonitrile in 10 mM ammonium bicarbonate pH 10 for 80
433 min, then they were pooled into multiple fractions (ten for label-free proteome) and
434 dried using vacuum centrifugation.

435

436 **Liquid chromatography-tandem MS analysis**

437 For label-free proteomic analysis, peptides were dissolved in 0.1% formic acid (FA,
438 94318-50ML-F, Fluka, USA) and loaded onto a reversed-phase pre-columns (Acclaim
439 PepMap 100, Thermo Scientific). Peptides were separated using a reversed-phase

analytical column (Acclaim PepMap RSLC, Thermo Scientific, USA). The elution gradient used was 5 % to 10 % of solvent B (0.1 % FA in 98 % ACN) for 4 min; 10 % to 23 % of solvent B for 36 min; 23 % to 35 % of solvent B for 12 min; 35 % to 80 % of solvent B for 4 min; and 80 % of solvent B for 4 min. Elution was performed at a constant flow rate of 300 nl/min using an EASY-nLC 1000 ultra-performance liquid chromatography (UPLC) system. Peptides subjected to post-translational modifications were separated in a similar way and analyzed using a Q Exactive TM Plus hybrid quadrupole-Orbitrap mass spectrometer (ThermoFisher Scientific, USA).

Peptides were subjected to a nanoelectrospray ionization source followed by tandem mass spectrometry (MS/MS) using a Q Exactive TM Plus (Thermo) mass spectrometer coupled online to a UPLC system. Intact peptides were detected using Orbitrap at a resolution of 70,000. Peptides were selected for MS/MS using a normalized collision energy of 30; ion fragments were detected using Orbitrap at a resolution of 17,500. A data-dependent procedure that alternated between one MS scan followed by 20 MS/MS scans was used for the top 20 precursor ions above a threshold ion count in the MS survey scan with different dynamic exclusion. For proteome detection, the ion count threshold and dynamic exclusion used were 1E4 and 30.0 s, respectively.

Analysis of global proteomics data

The resulting MS/MS data were processed using MaxQuant with an integrated Andromeda search engine (version 1.4.1.2). Tandem mass spectra were searched against a SwissProt human database (downloaded on August 2015) concatenated with a reverse decoy database. Trypsin/P was defined as the cleavage enzyme allowing up to two missing cleavages. For proteomic analysis, the first search range was set to 5 ppm for precursor ions, and the main search range was set to 5 ppm and 0.02 Da for fragment ions. The carbamidomethylation of cysteines was defined as the fixed modification, and the oxidation on methionine was defined as the variable modification. The quantification method used was LFQ, the FDR was adjusted to < 1 %, and the minimum score for modified peptides was > 40.

470

471 **Identification of DEPs**

472 To identify the proteins differentially expressed between tumor and healthy tissues,
473 DEPs were defined as meeting the following criteria: $|\log_2\text{-Fold change (FC)}| > 1$ and
474 $\text{FDR} < 0.05$. The heatmap of expression profiles was drawn using the *pheatmap*
475 package in R language. GO enrichment analysis was performed to assess the
476 functional biological role of DEPs. For GO term enrichment, including MF, biological
477 process, and cellular component, a two-tailed Fisher's exact test was used to
478 determine the enrichment of DEPs against all proteins identified in-house using a Perl
479 script. Categories with an $\text{FDR} < 0.05$ were considered significant. The graphical
480 summary of GO results was drawn using the *ggalluvial* R package.

481

482 **IPA**

483 The molecular and biological functions of DEPs were analyzed using IPA (Qiagen,
484 <https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/>) [44].
485 This analytical tool includes canonical pathway analysis, interaction network analysis,
486 disease and functional analysis, and a biomarker filter. The two statistical indicators of
487 IPA used were the p-value and z-score. P-values < 0.05 were considered statistically
488 significant, and the z-score was calculated using an internal algorithm and IPA
489 standard. The molecular interaction was activated when the z-score was > 0 and
490 inhibited when the z-score was < 0 .

491

492 **Screening and validation of diagnostic markers using ROC analysis**

493 The ROC curve was used to further select potential biomarkers with diagnostic power
494 by determining the specificity and sensitivity of each protein, and the AUC was used
495 to estimate the diagnostic value. The ROC curve was drawn using the *pROC* R
496 package.

497 For the testing cohort, a proteomic dataset from 60 GC patients obtained using
498 label-free analysis was used [16]. Raw data of all samples were downloaded, and the
499 same search engine MaxQuant as well as the same filter criteria were used to identify

DEPs. Based on these results, ROC analysis was conducted to test the diagnostic power of potential hub proteins in the testing cohort.

502

503 **Analysis using a random forest model**

504 To further build a multi-protein signature with diagnostic power, the significant hub
505 DEPs were used as attributes, and an analysis using a random forest model was
506 performed in-house using a python script. In brief, this method was used together with
507 10-fold cross-validation for the training set to build a four-protein signature. Moreover,
508 the diagnostic value of this model was verified using ROC analysis. Sensitivity,
509 specificity, accuracy, and AUC were used to determine predictive values.

510 In the testing phase, based on the expression of these four proteins, an analysis
511 using a random forest model was performed to prove the accuracy of the diagnostic
512 value of the protein signature.

513

514 **Validation of the four-protein signature using ELISA**

515 The diagnostic power of the four-protein signature was validated using 71 plasma
516 samples (37 from GC patients and 34 from healthy controls). In brief, human plasma
517 ATP synthase, H⁺ transporting, mitochondrial F1 complex, beta polypeptide (ATP5B),
518 ATP synthase, H⁺ transporting, mitochondrial F1 complex, O subunit (ATP5O),
519 NADH: ubiquinone oxidoreductase subunit B4 (NDUFB4), and NADH: ubiquinone
520 oxidoreductase subunit B8 (NDUFB8) levels were measured using commercial
521 ELISA kits (Catalog No. JL46945 for ATP5O, JL46944 for ATP5B, JL47136 for
522 NDUFB4, JL47139 for NDUFB8, Shanghai Jianglai Biotech, Shanghai, China)
523 according to the manufacturer's protocols.

524

525 **Validation of the four-protein signature using tissue microarray**

526 Immunohistochemistry was performed using 5 µm thick TMA sections as previously
527 described [45]. After dewaxing in xylene and rehydration through graded alcohol,
528 sections were placed in 3 % hydrogen peroxide to block endogenous peroxidase. For
529 antigen retrieval, sections were boiled in citrate/EDTA buffer (citrate buffer for

ATP5O, ATP5B and NDUFB4, Catalog No. ZLI-9065, ZSGB-BIO, Beijing, China;
 EDTA buffer for NDUFB8, Catalog No. R20904, Shanghai Yuanye Bio-Technology ,
 Shanghai , China). After blocking with goat serum (Catalog No. SP-9001 for ATP5B
 and NDUFB8, ZSGB-BIO, Beijing, China; Catalog No. SP-9002 for ATP5O and
 NDUFB4, ZSGB-BIO, Beijing, China), sections were incubated overnight with
 anti-ATP5O/anti-ATP5B/anti-NDUFB4/anti-NDUFB8 antibodies (anti-ATP5O:
 Catalog No. ab110276, diluted 1:5000, Abcam, Cambridge, UK; anti-ATP5B: Catalog
 No.33031, diluted 1:5000, SAB, College Park, Maryland, USA; anti-NDUFB4:
 Catalog No. ab110243, diluted 1:1000, Abcam, Cambridge, UK; anti-NDUFB8:
 Catalog No. ab192878, diluted 1:8000, Abcam, Cambridge, UK) at 4 °C. Sections
 were then further incubated with reagents from the SP IHC Kit (Catalog No. SP-9001
 for ATP5B and NDUFB8, ZSGB-BIO, Beijing, China; Catalog No. SP-9002 for
 ATP5O and NDUFB4, ZSGB-BIO, Beijing, China), 15 min for each reagent. Sections
 were visualized using DAB (Catalog No. ZLI-9010, ZSGB-BIO, Beijing, China) and
 Hematoxylin.

Slides were analysed with Panoramic MIDI digital scanner (3DHISTECH,
 Budapest, Hungary). Quantitative image analysis was performed by the Quant center
 software using histochemistry score (H-SCORE).

Statistical analysis

Statistical analyses were performed using the SPSS software version 19.0 (SPSS, Inc.,
 Chicago, IL, USA). The student's *t*-test was used to evaluate differences between DEP
 expression in GC tissues and in matched healthy tissues. ROC and AUC were used to
 assess the diagnostic value of each candidate protein and protein signature. Two-tailed
 p-values < 0.05 were considered statistically significant.

Data availability

557 The MS proteomics data were deposited in the iProX database at
558 <https://www.iprox.org/> with the dataset identifier IPX0001590000. The raw dataset is
559 free for all users.

560

561 Authors' contributions

562 YXS and JW designed experiments. JXS (Jingxu Sun), XWC, DHY and FZ
563 contributed to the literature review. JXS (Jinxin Shi) and ZHW collected the clinical
564 gastric cancer samples. YXS and JW helped to perform experiments. JW wrote the
565 initial draft of the manuscript. ZNW designed and supervised the study, developed the
566 concept and edited the paper. All authors have approved the final version of the
567 manuscript.

568

569 Competing interests

570 The authors declare no conflicts of interest.

571

572 Acknowledgements

573 We thank the department of Surgical Oncology of The First Hospital of China
574 Medical University for providing human gastric tissue samples and plasma samples.
575 We thank Jingjie PTM BioLab (Hangzhou, China) for technical assistance. We also
576 thank Elsevier (<https://webshop.elsevier.com/language-editing-services/>) for language
577 editing services.

578

579 Funding

580 This work was supported by the National Key R&D Program of China
581 (MOST-2017YFC0908300, MOST-2017YFC0908305), the National Science
582 Foundation of China (81872031), the Natural Science Foundation of Liaoning
583 Province (No. 20180550582) and the Project of Science and Technology of Shenyang
584 (18-014-4-07).

585

586 Reference

587 [1] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer
588 statistics, 2012. *CA Cancer J Clin* 2015;65:87-108.

589 [2] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in
590 China, 2015. *CA Cancer J Clin* 2016;66:115-32.

591 [3] Cancer Genome Atlas Research N. Comprehensive molecular characterization of
592 gastric adenocarcinoma. *Nature* 2014;513:202-9.

593 [4] Cui J, Yin Y, Ma Q, Wang G, Olman V, Zhang Y, et al. Comprehensive
594 characterization of the genomic alterations in human gastric cancer. *Int J Cancer*
595 2015;137:86-95.

596 [5] Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer.
597 *Lancet* 2016;388:2654-64.

598 [6] Marques-Lespier JM, Gonzalez-Pons M, Cruz-Correa M. Current Perspectives on
599 Gastric Cancer. *Gastroenterol Clin North Am* 2016;45:413-28.

600 [7] Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach:
601 A New Frontier in Cancer Research. *Biomed Res Int* 2018;2018:9836256.

602 [8] Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic
603 characterization of human colon and rectal cancer. *Nature* 2014;513:382-7.

604 [9] Latonen L, Nykter M, Visakorpi T. Proteomics of prostate cancer - revealing how
605 cancer cells master their messy genomes. *Oncoscience* 2018;5:216-7.

606 [10] Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al.
607 Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*
608 2016;534:55-62.

609 [11] Sharpnack MF, Ranbaduge N, Srivastava A, Cerciello F, Codreanu SG, Liebler
610 DC, et al. Proteogenomic Analysis of Surgically Resected Lung Adenocarcinoma. *J*
611 *Thorac Oncol* 2018;13:1519-29.

612 [12] Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated
613 Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*
614 2016;166:755-65.

615 [13] Zhang B, Whiteaker JR, Hoofnagle AN, Baird GS, Rodland KD, Paulovich AG.
616 Clinical potential of mass spectrometry-based proteogenomics. *Nat Rev Clin Oncol*
617 2018.

618 [14] Fernandez-Coto DL, Gil J, Hernandez A, Herrera-Goepfert R, Castro-Romero I,
619 Hernandez-Marquez E, et al. Quantitative proteomics reveals proteins involved in the
620 progression from non-cancerous lesions to gastric cancer. *J Proteomics*
621 2018;186:15-27.

622 [15] Jiang Z, Sun X, Zhang Q, Ji X, Yu Q, Huang T, et al. Identification of candidate
623 biomarkers that involved in the epigenetic transcriptional regulation for detection
624 gastric cancer by iTRAQ based quantitative proteomic analysis. *Clin Chim Acta*
625 2017;471:29-37.

626 [16] Ge S, Xia X, Ding C, Zhen B, Zhou Q, Feng J, et al. A proteomic landscape of
627 diffuse-type gastric cancer. *Nat Commun* 2018;9:1012.

628 [17] Cox J, Mann M. MaxQuant enables high peptide identification rates,
629 individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.
630 *Nat Biotechnol* 2008;26:1367-72.

631 [18] Srivastava A, Creek DJ. Discovery and Validation of Clinical Biomarkers of
632 Cancer: A Review Combining Metabolomics and Proteomics. *Proteomics*
633 2018:e1700448.

634 [19] Corbo C, Cevenini A, Salvatore F. Biomarker discovery by proteomics-based
635 approaches for early detection and personalized medicine in colorectal cancer.
636 *Proteomics Clin Appl* 2017;11.

637 [20] Chauvin A, Boisvert FM. Clinical Proteomics in Colorectal Cancer, a Promising
638 Tool for Improving Personalised Medicine. *Proteomes* 2018;6.

639 [21] Mikami T, Aoki M, Kimura T. The application of mass spectrometry to
640 proteomics and metabolomics in biomarker discovery and drug development. *Curr*
641 *Mol Pharmacol* 2012;5:301-16.

642 [22] Huang A, Zhang M, Li T, Qin X. Serum Proteomic Analysis by Tandem Mass
643 Tags (TMT) Based Quantitative Proteomics in Gastric Cancer Patients. *Clin Lab*
644 2018;64:855-66.

645 [23] Liu F, Zhang Y, Men T, Jiang X, Yang C, Li H, et al. Quantitative proteomic
646 analysis of gastric cancer tissue reveals novel proteins in platelet-derived growth
647 factor b signaling pathway. *Oncotarget* 2017;8:22059-75.

648 [24] Bostanci N, Selevsek N, Wolski W, Grossmann J, Bao K, Wahlander A, et al.
649 Targeted Proteomics Guided by Label-free Quantitative Proteome Analysis in Saliva
650 Reveal Transition Signatures from Health to Periodontal Disease. *Mol Cell*
651 *Proteomics* 2018;17:1392-409.

652 [25] Yang T, Xu P, Gu L, Xu Z, Ge W, Li Q, et al. Quantitative assessment of serum
653 heat shock protein 27 for the diagnosis of epithelial ovarian cancer using targeted
654 proteomics coupled with immunoaffinity enrichment. *Clin Chim Acta* 2018.

655 [26] Jiang Z, Zhang C, Gan L, Jia Y, Xiong Y, Chen Y, et al. iTRAQ Based
656 Quantitative Proteomics Approach Identifies Novel Diagnostic Biomarkers that were
657 Essential for Glutamine Metabolism and Redox Homeostasis for Gastric Cancer.
658 *Proteomics Clin Appl* 2018:e1800038.

659 [27] Vassilopoulos A, Fritz KS, Petersen DR, Gius D. The human sirtuin family:
660 evolutionary divergences and functions. *Hum Genomics* 2011;5:485-96.

661 [28] Frye RA. Phylogenetic classification of prokaryotic and eukaryotic Sir2-like
662 proteins. *Biochem Biophys Res Commun* 2000;273:793-8.

663 [29] O'Callaghan C, Vassilopoulos A. Sirtuins at the crossroads of stemness, aging,
664 and cancer. *Aging Cell* 2017;16:1208-18.

665 [30] Palmirotta R, Cives M, Della-Morte D, Capuani B, Lauro D, Guadagni F, et al.
666 Sirtuins and Cancer: Role in the Epithelial-Mesenchymal Transition. *Oxid Med Cell*
667 *Longev* 2016;2016:3031459.

668 [31] Li Z, Huang J, Shen S, Ding Z, Luo Q, Chen Z, et al. SIRT6 drives
669 epithelial-to-mesenchymal transition and metastasis in non-small cell lung cancer via
670 snail-dependent transrepression of KLF4. *J Exp Clin Cancer Res* 2018;37:323.

671 [32] Minn AJ, Wherry EJ. Combination Cancer Therapies with Immune Checkpoint
672 Blockade: Convergence on Interferon Signaling. *Cell* 2016;165:272-5.

673 [33] Park R, Williamson S, Kasi A, Saeed A. Immune Therapeutics in the Treatment
674 of Advanced Gastric and Esophageal Cancer. *Anticancer Res* 2018;38:5569-80.

675 [34] Bonotto M, Garattini SK, Basile D, Ongaro E, Fanotto V, Cattaneo M, et al.
676 Immunotherapy for gastric cancers: emerging role and future perspectives. *Expert Rev*
677 *Clin Pharmacol* 2017;10:609-19.

678 [35] Solaini G, Sgarbi G, Baracca A. Oxidative phosphorylation in cancer cells.
679 *Biochim Biophys Acta* 2011;1807:534-42.

680 [36] Warburg O. On respiratory impairment in cancer cells. *Science* 1956;124:269-70.

681 [37] Bruggemann M, Gromes A, Poss M, Schmidt D, Klumper N, Tolkach Y, et al.
682 Systematic Analysis of the Expression of the Mitochondrial ATP Synthase (Complex
683 V) Subunits in Clear Cell Renal Cell Carcinoma. *Transl Oncol* 2017;10:661-8.

684 [38] Xu G, Li JY. ATP5A1 and ATP5B are highly expressed in glioblastoma tumor
685 cells and endothelial cells of microvascular proliferation. *J Neurooncol*
686 2016;126:405-13.

687 [39] Sotgia F, Lisanti MP. Mitochondrial biomarkers predict tumor progression and
688 poor overall survival in gastric cancers: Companion diagnostics for personalized
689 medicine. *Oncotarget* 2017;8:67117-28.

690 [40] Ma J, Hou X, Li M, Ren H, Fang S, Wang X, et al. Genome-wide methylation
691 profiling reveals new biomarkers for prognosis prediction of glioblastoma. *J Cancer*
692 *Res Ther* 2015;11 Suppl 2:C212-5.

693 [41] Srinivasan S, Guha M, Kashina A, Avadhani NG. Mitochondrial dysfunction and
694 mitochondrial dynamics-The cancer connection. *Biochim Biophys Acta Bioenerg*
695 2017;1858:602-14.

696 [42] Hsu CC, Tseng LM, Lee HC. Role of mitochondrial dysfunction in cancer
697 progression. *Exp Biol Med (Maywood)* 2016;241:1281-95.

698 [43] Lleonart ME, Grodzicki R, Graifer DM, Lyakhovich A. Mitochondrial
699 dysfunction and potential anticancer therapy. *Med Res Rev* 2017;37:1275-98.

700 [44] Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in
701 Ingenuity Pathway Analysis. *Bioinformatics* 2014;30:523-30.

702 [45] Rizzardi AE, Zhang X, Vogel RI, Kolb S, Geybels MS, Leung YK, et al.
703 Quantitative comparison and reproducibility of pathologist scoring and digital image

704 analysis of estrogen receptor beta2 immunohistochemistry in prostate cancer. Diagn
705 Pathol 2016;11:63.
706

707 **Figure legends**

708 **Figure 1 Workflow of the study**

709 Screening of potential biomarkers for gastric cancer with diagnostic value using
710 label-free global proteome analysis.

711

712 **Figure 2 Proteomic features of differentially expressed proteins (DEPs) in** 713 **gastric cancer (GC)**

714 **A.** Heatmap of DEPs in GC. pT: pathological Tumor stage, pN: pathological Lymph
715 Node stage, pTNM: pathological TNM stage. **B.** A volcano plot for DEPs. The
716 differential expression ratio of $\log_2$ (FC) (x axis) and the $-\log_{10}$ (FDR) value (y axis)
717 were plotted for each identified protein. **C.** Expression profiles of the top ten
718 significant DEPs in tumor and normal tissues. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$
719 (paired-samples *t*-test).

720

721 **Figure 3 Subcellular distribution and GO term enrichment analysis of** 722 **differentially expressed proteins (DEPs) in gastric cancer (GC)**

723 **A.** Subcellular distribution of DEPs. **B.** Graphical summary of DEPs and GO term
724 enrichment analysis.

725

726 **Figure 4 Activated or inhibited pathways and potential hub proteins in gastric** 727 **cancer (GC) using ingenuity pathway analysis**

728 **A.** Canonical pathway analysis of DEPs. **B.** Expression profiles of proteins in the
729 oxidative phosphorylation pathway and the sirtuin signaling pathway. **C.** Disease and
730 functional analysis for DEPs.

731

732 **Figure 5 Interaction network and biomarker analysis for DEPs**

733 **A.** Interaction network analysis of tissue invasion by tumor cell lines and lipid export.
734 **B.** Biomarker analysis for DEPs.

735

Figure 6 Screening and validation of potential proteins as diagnostic markers for gastric cancer in our and testing datasets

A. ROC curve for each hub protein in the training phase. **B.** Venn diagram summarizing the number of DEPs between our study and the testing phase. **C.** ROC curve for each shared hub protein in the testing phase. **D.** Identification of potential and independent diagnostic proteins as biomarkers. **E.** Expression profile of significant shared hub proteins in the testing phase.

Figure 7 Establishment and validation of a four-protein signature for the diagnosis of gastric cancer

A. Design of the random forest model. **B.** Feature performance of each protein in the random forest model. **C.** Increment feature selection. AUC, sensitivity, specificity, and accuracy became constant and suitable when the four-protein signature was built. Sen: sensitivity, Spe: specificity, Acc: accuracy. **D.** ROC curve for the four-protein diagnostic signature for the training set (red line) and testing set (blue line). **E.** ROC curve for the four-protein diagnostic signature for the plasma samples.

Figure 8 Validation of a four-protein signature for the diagnosis of gastric cancer in tissue microarray

The immunohistochemical staining of **A.** ATP5B, **B.** ATP5O, **C.** NDUFB4, and **D.** NDUFB8 in tumor and normal GC tissues. Scale bars represent 50 and 100 μ m. Each case was at $\times 200$ magnification. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$ (paired-samples t -test). **E.** ROC curve for the four-protein diagnostic signature for the tissue microarray.

Supplementary Material

Supplementary Tables Legends

Supplementary Table 1 The clinical information across all AGC samples in this study.

765 **Supplementary Table 2** Filtered proteins used in further analysis.

766 **Supplementary Table 3** Identification of differentially expressed proteins in AGC
767 by proteomic analysis.

768 **Supplementary Table 4** GO term enrichments analysis of DEPs.

769 **Supplementary Table 5** IPA canonical pathway analysis of DEPs.

770 **Supplementary Table 6** 15 potential hub proteins by ingenuity pathway analysis.

771 **Supplementary Table 7** The 95% confidence interval (CI) of AUC and p-value
772 compared with reference line in this study and Qin's study.

773 **Supplementary Table 8** Identification of differentially expressed proteins in testing
774 phase (Qin's study).

775 **Supplementary Table 9** The p-value compared with reference line in all datasets.

776

777 **Supplementary Figure Legends**

778 **Supplementary Figure 1 Quality control of clinical samples and global
779 proteome profiling**

780 **A.** The neoplastic purity analysis of all 60 samples. **B.** The distribution of peptides,
781 unique peptides, and identified proteins in this study.

782

783 **Supplementary Figure 2 General overview of global proteome profiling**

784 **A.** Proteomic data filtered at different levels. FDR: false discovery rate, CV:
785 coefficients of variation, IQR: interquartile range, FOT: fraction of total. **B.**
786 Distribution of the protein sequence coverage across all 60 samples from 60 gastric
787 cancer patients. **C.** Distribution of quantified proteins across all 60 samples. **D.**
788 Spearman's correlation coefficient of the protein ratios in 30 GC tissues and 30
789 matched healthy tissues. **E.** Distribution of $\log_{10}(\text{FOT})+7$ of the identified proteins in
790 30 GC tissues and 30 matched healthy tissues. Boxes range from the lower quartile to
791 the upper quartile.

792

793 **Supplementary Figure 3 General overview of proteome quantification in AGC**

794 **A.** Relationship between coefficient of variance (CV) and $\log_{10}(\text{FOT})+7$. When FOT
 795 reached more than 10^{-5} , overall CV significantly decreased. **B.** Relationship between
 796 interquartile range (IQR) and $\log_{10}(\text{FOT})+7$. The increased performance of IQR was
 797 discontinued after the FOT reached $> 10^{-5}$. **C.** Relationship between quantitative
 798 samples and median CV. When quantitative samples reached > 20 , median CV
 799 significantly decreased. **D.** PCA analysis showing tumor and matched healthy tissues.
 800 PC: principal component.

801

802 **Supplementary Figure 4 The AUC values of each protein from the four-protein**
 803 **combination in these datasets**

804 **A.** The AUC of each protein in Qin's study (LC-MS/MS). **B.** The AUC of each
 805 protein in plasma (ELISA). **C.** The AUC of each protein in tissue microarray.

Table 1 The information of top 10 significant DEPs

UniProt ID	Gene Symbol	T-test p-value	FDR	Log2 T/N Ratio	Regulation
P20648	ATP4A	5.15E-09	1.02E-07	-4.66	down
P51164	ATP4B	1.53E-07	1.24E-06	-4.37	down
Q96NY7	CLIC6	7.13E-10	2.55E-08	-4.16	down
P12277	KCRB	1.38E-10	8.61E-09	-4.12	down
P07098	LIPG	2.56E-15	1.01E-11	-3.70	down
Q9UGM3	DMBT1	3.10E-02	4.48E-02	4.92	up
P36952	SPB5	3.78E-05	1.13E-04	3.92	up
P31327	CPSM	1.33E-02	2.09E-02	3.62	up
P46013	KI67	2.39E-03	4.39E-03	3.05	up
P06731	CEAM5	3.76E-04	8.45E-04	2.71	up

Cohort 1-Training set

Samples collection

30 pairs of advanced gastric cancer and normal clinical samples

Label-free global proteome analysis

Further data analysis

Protein signature

Four protein-signature

ATP5B-ATP5O-NDUFB4-NDUFB8

Cohort 2-Testing set

Testing data set

Proteomic landscape of 60 pairs of gastric cancer from Qin et al. study

The same search engine

MaxQuant

Differentially expressed proteins screening

ROC of candidate proteins as biomarkers

Diagnostic biomarkers

Cohort 3-Two independent sets

Blood plasma samples

ELISA of plasma in gastric cancer and healthy controls

Tissue microarray

A

B

C

A

B

A

B

C

Predicted relationships

- Leads to activation
- Leads to inhibition
- Findings inconsistent with state of downstream molecule
- Effect not predicted

A

B

This study

Qin study

C

D

E

A

B

C

D

E

A

ATP5B

B

ATP5O

C

NDUFB4

D

NDUFB8

E