CagA sequence differences and proteome profiles between East Asian and Western *H. pylori* strains

Jianjiang Zhou (zjj1985048@gmc.edu.cn)
Guizhou Medical University

Yan Zhao
Guizhou Medical University

Yang Xie
Guizhou Medical University

Lin Xiong
Guizhou Medical University

Xinying Quan
Guizhou Medical University

Qin-Rong Wang
Guizhou Medical University

Wen-Ling Wang
Guizhou Medical University

Feng Gin
Guizhou Medical University

Qi-Fang Zhang
Guizhou Medical University

De-Zhong Liao
Guizhou Medical University

Lucas Zellmer
University of Minnesota Cancer Center: Masonic Cancer Center

Research article

Keywords: Helicobacter pylori, CagA, East Asian strains, Western strains, proteomics

Posted Date: November 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-112813/v1

Abstract

Background

The cytotoxin-associated gene A protein (CagA), an effector protein of *Helicobacter pylori (H. pylori)*, was the first identified bacterium oncoprotein. Based on its sequence characteristics, *H. pylori* has been classified into East Asian and Western strains. We hypothesized that the differences in structure of CagA and proteomic profiles between East Asian and Western *H. pylori* strains are the primary cause of the differential clinical outcomes of *H. pylori* infection.

Results

In this study, we isolated 27 *H. pylori* strains from gastric mucosa of Chinese patients with gastric diseases and revealed that the Western CagA has more variation in its EPIYA motifs than East Asian CagA. This result was further confirmed via analyzing the CagA sequences of 150 *H. pylori* strains from GenBank. More importantly, we detected the deletion or partial deletion of 13 amino acids in CagA in all East Asian strains but not in Western strains. iTRAQ -based proteomic analysis showed that the CagA protein related to cytotoxicity was highly expressed in Western strains, while urease-associated proteins UreA and UreH, flagellin related proteins FlaA, FlgE and FlhA, and cell division proteins FtsZ and FtsI were highly expressed in East Asian strains. These proteins were associated with the colonization, motility, and viability potential of *H. pylori* in the human stomach and were clustered into an interaction network.

Conclusions

This study provides significant, unreported differential sequences of CagA and proteomic profiles between East Asian and Western *H. pylori*, which maybe serves as promising new targets to ascertain the pathogenesis of *H. pylori*.

Background

Helicobacter pylori (H. pylori), a spiral-shaped, flagellated, microaerophilic Gram-negative bacillus infecting over 50% of the world's population, was classified as a Group I human carcinogen by an International Agency for Research on Cancer (IARC) Working Group in 1994, and was reconfirmed by a new Working Group with sufficient evidence of causing non-cardia gastric carcinoma via chronic infection with *H. pylori* in 2009 [1]. The experimental studies in Mongolian gerbils also provided the strong evidence for the carcinogenicity of infection with *H. pylori* [2]. Although the prevalence of *H. pylori* infection appears to be decreased in certain parts of the world, it remains high in many regions and countries, particularly in African and Asian countries. It has recently been estimated that 79% of new gastric cancers diagnosed each year globally can be attributed to *H. pylori* infection, making it a primary risk factor for gastric cancer [3, 4].

Gastric cancer is the sixth most common malignancy and the fifth most common cause of cancer death worldwide based on updated estimates from IARC [5, 6]. However, gastric cancer incidence and mortality vary significantly in the different countries and regions. According to IARC in 2018, Asia had the highest incidence and mortality of gastric cancer with 14.3 and 10.7 ASR (age-standardized rate, world) per 100,000, which is 3.49 and 5.94 times than that of the lowest North America with 4.1 and 1.8 ASR per 100,000, respectively. Accordingly, a meta-analysis showed that Asian regions had an *H. pylori* prevalence of 54.7%, higher than the prevalence of 37.1% in North American in 2017 [7]. Another systematic review concluded that the prevalence of *H. pylori* infection was at least two-fold higher in countries with high gastric cancer incidence compared with countries with a lower gastric cancer rate [8]. These data suggested the close association between *H. pylori* infection and gastric cancer development.

CagA, a well-characterized effector protein of *H. pylori*, is encoded by the *cagA* gene located in the cag pathogenicity island (cag PAI) of the pathogen genome. This protein is injected into gastric epithelial cells via a type IV secretion system after infection and colonization of the human stomach with *H. pylori* [9]. The translocated CagA in the host cells, as a bacterial oncoprotein, is phosphorylated on tyrosine residues of Glu-Pro-Ile-Tyr-Ala (EPIYA) motifs of this protein by host tyrosine kinases of the Src and Abl families [10]. The phosphorylated CagA mimics the proteins of the host cell to interact with numerous signaling factors and hijack intracellular signaling transduction pathways, contributing for chronic gastric disorders and gastric cancer [11]. A meta-analysis investigating CagA seropositivity irrespective of *H. pylori* status concluded that the individuals with *H. pylori* infection have a higher risk of gastric cancer if their serum CagA is positive than those with serum CagA negative [12].

EPIYA and CM (CagA multimerization) segments are two critical motifs of CagA. Based on the amino acid sequences flanking each of the EPIYA motifs, four distinct EPIYA segments (EPIYA-A, -B, -C and -D) have been defined and thereby CagA was classified as Western CagA and East Asian CagA [13]. Western CagA is comprised of EPIYA-A, -B and single or multiple -C motifs while East Asian CagA possesses EPIYA-A, -B, and -D segments in tandem [14]. Clinical epidemiological studies showed that gastric cancer is more closely associated with East Asian CagA than Western CagA [15]. Experimental studies also demonstrated that East Asian CagA has a higher affinity with the pro-oncogenic SHP-2 phosphatase, a downstream effector of CagA, than Western CagA [16] Approximately 60% of Western *H. pylori* stains are *cagA*-positive and patients infected with this strain develop cancer at higher rates in Western populations [17, 18]. However, in East Asian countries, almost all *H. pylori* strains possess CagA and individuals infected with the *cagA*-positive *H. pylori* have a closer association with gastric mucosal atrophy, an important precancerous lesion of gastric cancer [17]. However, in recent years, the carcinogenic effects of CagA have been questioned, especially in Asian countries, since only about 1% of individuals infected with *cagA*-positive *H. pylori* progress to gastric cancer while the underlying mechanism is not understood [19].

The proteome studies on *H. pylori* and infected hosts have been carried out in several laboratories to investigate the protein phosphorylation, protein-protein interaction, and strain-level typing of *H. pylori* [20]. Using differential proteome analysis, Müller et al found outer membrane proteins were more abundant,

whereas proteins involved in cell division, transcriptional and translational processes, and host colonization were less abundant in the coccoid forms of *H. pylori* compared with the infectious spiral form [21]. However, the proteome profiles between East Asian and Western *H. pylori* strains have not been documented to date.

Taken together, East Asian *H. pylori* has a stronger potential to induce chronic atrophic gastritis and gastric cancer than Western strains. However, *H. pylori*-CagA status alone cannot entirely reasonably explain the different clinical outcomes. Other proteins encoded by the *H. pylori* genome may be associated with this difference. Therefore, in this study, we isolated 27 clinical strains of *H. pylori* from the gastric mucosa of Chinese patients with gastric disease and compared *H. pylori*-CagA sequences of these strains. Subsequently, isobaric tags for relative and absolute quantitation (iTRAQ) coupled LC-MS/MS-based quantitative proteomics was used to obtain the differentially expressed proteome between East Asian and Western *H. pylori* stains. In conclusion, we found the unreported differences of *H. pylori*-CagA sequences and obtained the differential proteomic pattern between East Asian and Western *H. pylori* strains for the first time.

Results

Identification and characteristics of clinical *H. pylori* isolates

Twenty-seven clinical strains of *H. pylori* were successfully isolated from the gastric mucosa of Chinese patients with chronic gastritis, atrophic gastritis, gastric ulcer, and gastric cancer and identified by gram staining, urease, catalase, and oxidase tests. Sequencing of the *16S rDNA* and *cagA* genes of *H. pylori* was performed and showed all of the 27 isolates were *cagA*-positive (Additional file 1: Fig. S1). Subsequently, the *cagA* gene sequences amplified from 30 *H. pylori* strains by PCR, including 26 sequences originating from clinical isolates named GZ1-GZ26 and four sequences from Western strains NCTC11637, NCTC11639, 26695, and SS1, were submitted to the GenBank database with accession numbers KR154731-KR154758, GQ161098, GQ161099, KR154758, and KR154757. One strain (GZ15) was not submitted to the GenBank database due to the loss of the C-terminal region of CagA.

Based on the characteristics of EPIYA motifs of CagA, 20 of the 30 strains were classified as East Asian strains with the remaining 10 strains as Western strains, including six clinical isolates (Table 1). In addition, two East Asian strains (GZ17 and GZ18) were isolated from the same patient and East Asian strain GZ11 and Western strain GZ10 were isolated another patient. Strain GZ15 lacks EPIYA-A, -B, and -C/D sites and cannot be divided into any group.

The sequence comparison of CagA showed that Western strains had more variation than East Asian strains in EPYIA motifs, especially with the deletion of the EPIYA-C site (5/10 strains) and variation of A→T at the EPIYA-B site (5/10 strains), while only 2 of the 21 East Asian strains had A→V variation at the EPIYA-D site and one strain showed P→S conversion at the EPIYA-B site (Table 2). More importantly, the deletion or partial deletion of 13 amino acids at the 821^{th}-834^{th} region of CagA (referring to 26695

sequences) was detected for the first time in all East Asian strains but not in Western strains (Fig. 1A). An eight amino acid difference between the EPIYA-C/D side and CM motifs between East Asian and Western strains was also detected (Fig. 1B).

Functional domain diversity of CagA between East Asian and Western strains

Studies on the crystal structure of CagA revealed that CagA consists of an N-terminal ordered region (residues 1-829), including domain I, II, III, and a C-terminal disordered segment (residues 830-1186) [11]. There are some important functional domains and segments in CagA, including the CagA-phosphatidylserine (PS) interaction domain, EPIYA motifs, and CM motifs (Fig. 2A). The PS segment in domain II mediates the attachment of CagA to the cytoplasmic membrane past-translocation, while N- and C-terminal binding sequences (NBS and CBS) are associated with CagA dimerization. EPIYA motifs (Glu-Pro-Ile-Tyr-Ala) have several important functions, such as CagA phosphorylation, binding of CagA to SHP-2, and activation of multiple intracellular signaling pathways. CM motifs, by binding to partitional defective 1 kinase b (PAR1b, a serine-threonine kinase), participate in the maintenance of gastric epithelial cell polarity. In brief, these domains play a crucial role in *H. pylori*-induced gastric pathogenesis. Interestingly, we found that there is significant sequence variability in these important domains and their flanking regions between East Asian and Western CagA (Fig. 2B).

Construction of molecular phylogenic tree of CagA

One hundred and fifty sequences of *H. pylori*-CagA, including our own isolates, were randomly obtained from the Genbank database to construct the CagA-based molecular phylogenic tree via MEGA software (Fig. 3). In the phylogenic tree, a total of 150 strains were clustered into two large groups: East Asian group with characteristics of EPIYA-ABD (80 strains) and Western group with characteristics of EPIYA-ABC (70 strains). Western group was further clustered into three subgroups that was named as Western group, East Asian type of the Western group, and South America type of the Western group based on its geographical origin. The East Asian group primarily originated from China (20/80), Japan (17/80), Vietnam (29/80), and the Philippines (6/80). The Western Group primarily originated from Colombia (24/70), and Philippines (12/70). However, many strains originating from East Asian countries including Japan, China, and Korea are clustered into the East Asian type of the Western group; two strains originating from Peru are clustered into the South America type of the Western group. In 24 of our clinical isolates, 18 and six isolates are clustered into East Asian and Western groups, respectively, suggesting that Western *H. pylori* strains are widespread in China. We further found that strains from the same country and region tend to cluster together, which may be the cause of varying gastric cancer incidence in different geographical regions. According to global estimates of the incidence and mortality rates by the World Health Organization in 2018, gastric cancer incidence gradually decreases from East Asian regions like Korea, China, and Japan to South-East Asian regions, such as Vietnam and the Philippines, and finally to South American regions, such as Colombia.

The sequences flanking EPIYA-C/D sites affect tyrosine phosphorylation of CagA and are defined as left and right CM domains, respectively. We found three sets of different sequences in both sides of EPIYA-

C/D sites. The FLLKRHDKVDDLSKVG is a typical sequence located on both sides of EPIYA-C in the Western group and East Asian type of the Western group; it represents the classical CM motifs. The SSLKRYAKVDDLSKVG is located on both sides of EPIYA-C in the South America type of the Western group, which was reported as less virulent strains. The third set of sequences is found on both sites of EPIYA-D in the East Asian group. The KIASAGKGVGGFSGVG segment to the left of EPIYA-D replaces the classical CM motif on the left of EPIYA-C and the FPLRRSAAVND LSKVG to the right of EPIYA-D is partly identical to EPIYA-C (Additional file 2: Fig. S2). In addition, consistent with the results analyzed from our isolates, Western CagA has more variation, in which conversion of EPIY**A** to EPIY**T** at the EPIYA-B site, and duplication of EPIYA-C, appear in 25 (36%) and 31 (44%) of the 70 Western strains, respectively, while the same change at the EPIYA-B site was observed only in 4 (5%) of the 80 East Asian strains. The variations in EPIYA motifs of CagA from 150 strains was listed in Table 2.

We also analyzed the composition of 20 amino acids in East Asian and Western CagA from 150 *H. pylori* stains and found that certain amino acids, such as Glu, Leu, Thr, Arg, and Trp are significantly more present in East Asian strains. Conversely, Lys, Met, Gly, His, Pro, Val, and Cys are significantly more present in Western strains (Additional file 3: Fig. S3). In particular, 50/70 Western CagA contain Cys, an important amino acid in protein dimerization, but only 12/80 East Asian CagA has Cys.

iTRAQ-based quantitative proteomics of East Asian and Western *H. pylori* strains

CagA sequence polymorphisms between East Asian and Western *H. pylori* strains cannot completely explain their pathogenic differences. Therefore, we sought to further define the proteomic changes of two strain groups. Six *H. pylori* strains, including three East Asian strains (GZ1, GZ3, and GZ7) with EPIYA-ABD motifs and three Western strains (NCTC11639, 26695, and GZ5) with EPIYA-ABC motifs) were used to conduct iTRAQ-based absolute quantitative proteomics. Proteomic analysis quantified a total of 2084 proteins and 108 differentially expressed proteins between Western and East Asian *H. pylori* strains according to the criteria of Bonferroni-corrected $P < 0.01$, and fold change ≥ 1.2 (up-regulation) or ≤ 0.8 (down-regulation) [22]. After exclusion of the hypothetical, duplicate, and unidentified proteins, 70 differential proteins are mapped to the standard strain *H. pylori* 26695 in Uniprot Database (https://www.uniprot.org/), of which 26 proteins were up-regulated and 44 proteins were down-regulated in the Western group compared to the East Asian group (Fig. 4A and 4B). Using hierarchical clustering analysis, we found that, among differential proteins, CagA protein was highly expressed in the Western group. Alternatively, the urease subunit alpha (UreA) and urease accessory protein (UreH), both of which are related to the colonization of *H. pylori* in the human stomach, were highly expressed in the East Asian group. We further observed that flagellin-associated proteins (flagellin FlaA, flagellar hook protein FlgE, and flagellar biosynthesis protein FlhA) and cell division proteins (FtsZ and FtsI) were found in abundance in the East Asian group (Fig. 4C). More details of differential proteins are presented in Additional file 4: Table S1, and the identification of differential proteins by MS/MS are presented in Additional file 5:Table S2.

Because of the high heterogeneity among different *H. pylori* strains, the biological repeatability of three strains in same group was analyzed by Principal Component Analysis (PCA) and correlation coefficients of normalized protein intensity between two strains of the same group were measured. The results indicated good clustering and clear distinction for both groups (Fig. 4D). The correlation coefficient ranges from 0.45 to 0.63 (Additional file 6: Fig. S4). The standard deviation and coefficient of variation of the abundance of 2084 proteins and 70 differential proteins were also calculated and shown in Additional file 7: Table. S3 and Additional file 8: Table. S4, respectively. Finally, the mRNA expression levels of five differentially expressed proteins including UreA, FlaA, FlgE, CagA, and FlhA were validated by RT-qPCR and the validation results were consistent with the proteomic results (Fig. 4E).

Functional annotation and protein interaction networks of differential proteins

To obtain functional information and interaction networks, 70 differentially expressed proteins were annotated with gene ontology (GO) by DAVID 6.8. KEGG pathway enrichment and interaction network analysis were conducted by KOBAS 3.0 and STRING online tools, respectively. The results showed that these differential proteins are mainly associated with biosynthetic processes, metabolism, translation and gene expression (Fig. 5A), and are enriched into nine important pathways, in which five pathways possess significant enrichment (FDR-corrected P value < 0.05) (Fig. 5B and 5C). Protein-protein interaction analysis indicated that the highly-expressed proteins in East Asian strains are clustered into two significant networks with UreA and FtsI as core nodes, while the highly-expressed proteins in Western strains are clustered into an important network with GroEL and CagA as nodes (Fig. 5D).

Discussion

Evidence including epidemiological, clinical, and experimental studies, transgenic models, and *H. pylori* infection of Mongolian gerbils concluded that chronic infection with *H. pylori cagA*-positive strains is the strongest risk factor of gastric cancer [11, 23–24]. A meta-analysis also determined that eradication of *H. pylori* is associated with a reduction of gastric cancer risk [25]. However, *H. pylori* prevalence rate, virulent strains, and gastric cancer incidence are highly variable in different countries and regions throughout the world. The incidence of gastric cancer in East Asian countries such as Japan, China, and Korea is almost ten-times higher than that in the United States [26].

CagA is the first bacterial oncoprotein identified in human cancer and its sequences determine the classification and geographical origin of *H. pylori* strains. In this study, we found that 23% (6/26) strains isolated from the gastric mucosa of Chinese patients are Western strains, suggesting that Western *H. pylori* is prevalent in China. We also detected the co-infection of East Asian and Western strains in one individual. Moreover, we found that all isolates of both East Asian and Western strains are *cagA*-positive. This result is consistent with another report in which *H. pylori* strains isolated from East Asian countries have a higher *cagA*-positive rate (90%~100%) than strains from a Western population (~ 60%) [27].

EPIYA motifs of CagA are an important tyrosine-phosphorylation (pY) site. Accumulated evidence has showed that the sequence polymorphism and variable alignments of EPIYA segments has been linked to

the pathobiological action of individual CagA [26]. We found that 5 of the 10 Western strains in this study have a deletion of the EPIYA-C site and the other five strains have a variation of A→T at the EPIYA-B site. However, in 20 East Asian strains, only two strains have a A→V conversion at the EPIYA-D site, one strain has a P→S conversion at the EPYIA-B site, and one strain lacks the EPIYA-A site. The data suggest that the EPIYA motifs of Western strains have more variation than East Asian strains and the variation pattern of the two groups of strains is different. The results are further confirmed by the phylogenic trees, based on 150 *H. pylori*-CagA amino acid sequences from the GenBank database, in which 36% (25/70 strains) and 44% (31/70 strains) of Western strains carry the variation of A→T conversion at the EPIYA-B site and duplication of EPIYA-C, respectively. Conversely, A→T conversion of the EPIYA-B site was observed in only 5% (4/80 strains) of East Asian strains. The A/T polymorphism in the EPIYA-B motif was reported to influences the function of CagA, in which CagA with an EPIY**T**-B motif has a higher affinity with PI3K compared to CagA with an EPIY**A**-B motif, subsequently leading to an increased secretion of IL-8 [28]. CagA with multiple EPIYA-C motifs was also confirmed to have a stronger pro-carcinogenetic potential [28]. Although the deletion of EPIYA-C was less documented,our results were confirmed by two Western strains NCTC 11639 (AB015416) and IND 07 (LC339379) submitted into GenBank from Japan and Indonesia, respectively, that have the same loss of EPIYA-C. The EPIYA-D motif polymorphism of East Asian CagA is rarely reported. We detected the A→V conversion at the EPYIA-D site in two clinical isolates which have not been found in any other *H. pylori* strains, including Western and East Asian strains. More importantly, we, for the first time, identified the deletion or partial deletion of 13 amino acids downstream to the N-terminal binding sequence of CagA in all East Asian *H. pylori* studied.

Apart from the above-mentioned sequence differences between Western and East Asian strains, other unreported variations were also detected in this study, such as CM motifs. We note that the eight amino acid difference located between EPIYA-C/D and the right CM motif among the two kinds of strains was also unexplored. We also found the differences in amino acid composition of CagA between the two groups of strains. These findings suggested that there are potentially more valuable differences between East Asian and Western strains that demand further investigation.

The proteome difference between East Asian and Western *H. pylori* strains is not reported. Through iTRAQ-based absolute quantitative proteomic analysis, we quantified a total of 2084 proteins and identified 70 differentially expressed proteins with functional annotation between East Asian and Western strains; 26 of these proteins were up-regulated and 44 proteins were down-regulated in Western strains compared to East Asian strains. Importantly, the comparison of proteome profiles indicated that CagA protein was more abundant in Western strains. This finding may partly explain the observed difference that in Western populations, *cagA*-positive strains are associated with enhanced induction of gastritis, gastric ulcers, and a higher risk of gastric cancer. However, in East Asian populations where almost all strains are *cagA*-positive, the *cagA* gene is not associated with an increased risk of gastric diseases [29].

Our study is the first to report that the proteins involved in host colonization (UreA and UreH), cell division (FtsZ and FtsI), and cell movement (flagellin FlaA, flagellar hook protein FlgE, and flagellar biosynthesis protein FlhA) were more abundant in East Asian strains. Among these proteins, FlaA allows *H. pylori* to

migrate into the host gastric epithelium to survive in a comparatively higher pH niche since *H. pylori* is not an acidophile [30, 31]. FlhA and FlaE are reported to help flagellar biosynthesis and functioning [32]. UreA matures to form an active enzyme by combining with nickel (Ni) to transform urea to ammonia with the help of Ni binding protein hypA, FtsZ and UreH [33]. The ammonia can decrease the pH of the gastric mucosal layer, ultimately promoting the colonization and persistent infection of *H. pylori*. A recent study revealed that urea emanating from the gastric epithelium can attract *H. pylori* by binding to chemotaxis protein TlpB on the bacterium's surface in the presence of a powerful urease and direct the bacterium's movement toward the gastric epithelium [34]. To verify the proteomic results, we detected the mRNA levels of the UreA, FlaA, FlgE, CagA, and FlhA by RT-qPCR because of the lack of the suitable antibodies and obtained the consistent results, suggesting that the East Asian strains of *H. pylori* may have stronger colonization and mobility capability in the human stomach compared to Western strains.

Further analysis indicated that the 70 differential proteins are mapped to five significant pathways, including microbial metabolism in diverse environments, glyoxylate and dicarboxylate metabolism, and DNA replication. GO enrichment analysis showed that these proteins were primarily associated with the biosynthetic process, metabolism, translation, and gene expression of *H. pylori*. More importantly, utilizing proteomic data, we found the proteins highly expressed in East Asian strains were enriched into two key PPIs with UreA and FtsI as core nodes, while the proteins highly expressed in Western strains were enriched into one important PPI with CagA and GroEL as nodes. However, more experiments are required to confirm these results.

To analyze the influence of the in vitro passage numbers of *H.pylori* on proteome expression in bacteria, the PCA, correlation coefficient of two strains, standard deviation (SD), and coefficient of variation (CV) of original data set were computed. PCA results indicated good clustering and clear distinction between East Asian and Western strain. The correlation coefficient between low-passage-number clinical isolate (GZ5) and high-passage-number strains (11639 or 26695) was higher than that between two high passage 11639 and 26695 strains in Western group. SD value of the abundance of 2084 proteins from three East Asian strains GZ1,GZ3, and GZ7 was from 0.0092 to 4.3676, while this value from Western strains GZ5, 11639, and 26695 was from 0.0047 to 3.9080. Similarly, 98% of 2084 proteins in both East Asian and Western groups had a CV value less than 10%. These results indicated good reproducibility of the methods and less dispersion of data set, which suggested that effects of passage in vitro on protein expression of *H.pylori* is small in our study.

Conclusions

This study compared the amino acid sequences of CagA and proteomic profile between East Asian and Western *H. pylori* strains and found many significant and unreported differences. The results provide significant evidence of new differential sequences of CagA and proteomic profiles between East Asian and Western strains, which maybe serves as new study targets to determine the pathogenesis of *H. pylori* and to elucidate the mechanism underlying the development and progression of *H.pylor*i-induced gastric diseases.

Methods

H. pylori strains

H. pylori 26695 (ATCC 700392) were purchased from American Type Culture Collection (ATCC). *H. pylori* NCTC 11637 (ATCC 43504), NCTC 11639 (ATCC 43629), and SS1 were gifts from the Chinese Center of *H. pylori* strain Management and Preservation, Beijing, China. The four Western strains were used from 3rd to 5th passage in our laboratory after receipt. All of the *H. pylori* strains are *cagA*-positive and were grown on Columbia agar plate containing 10% sheep blood and *H. pylori* selective supplement (Oxoid Ltd, England) at 37°C under microaerobic conditions.

Isolation, culture, and identification of clinical *H. pylori* isolates

All of the samples were collected under endoscopy from gastric mucosa of patients with gastric diseases from January to December 2017 in the Affiliated Hospital of GuiZhou Medical University, China. The samples, within four hours of excision, were fully cut into pieces, homogenized, and then inoculated on Colombian agar plate containing 10% sheep blood and *H. pylori* selective supplement. After bacterial culture for 3-5 days at 37°C under microaerobic conditions, a single colony was further isolated, purified, and cultured. The isolates successfully obtained were identified by gram staining, urease, catalase, and oxidase testing and sequencing of the *16S rDNA* of *H. pylori* (sense primer: 5'-CTTGCTAGAGTGCTGATTA-3', antisense primer: 5'-TCCCACACTCTAGAATAGT-3'). The protocol was approved by the Ethics Committee of Guiyang Medical University and all subjects provided written informed consent. The identified isolates were used from 3rd to 5th passage after purification.

Acquisition of full-length sequences of *H. pyloricagA*.

DNA was extracted from identified *H. pylori* strains. The full-length *cagA* sequence was synthesized by PCR (sense primer: 5'-AACAATGACTAACGAAACCA-3', antisense primer: 5'-TAAAGAA TGGCTCAAATTGT-3', about 4000 bp) and cloned into pMD18-T plasmids to construct pMD18-T/*cagA* vector, which was identified by sequencing. The *cagA* sequences successfully obtained were submitted to the GenBank database.

Sequence alignment and clustering analysis of CagA

Sequences of the *cagA* gene were converted into amino acid sequences by DNAstar software and multi-sequence alignment and cluster analysis were performed using Clustal Omega software based on the amino acid sequences of CagA.

Next, 150 amino acid sequences of *H. pylori* CagA including our isolates were randomly obtained from the Genbank database. MEGA 4.0 software was used to construct CagA-based molecular phylogenetic trees with Neighbor-Joining and *P*-distance methods.

Isobaric tags for relative and absolute quantitation (iTRAQ)

Six *H. pylori* strains, including three East Asian strains and three Western strains, were taken out from liquid nitrogen stocks at same time and resuscitated via culturing them for 3-5 days on Columbia agar plate containing 10% sheep blood and *H. pylori* selective supplement. Then these bacteria were cultured overnight in liquid medium under microaerobic conditions and harvested by centrifugation. Six samples were ground in liquid nitrogen. The bacterial protein was extracted by ultrasound method (ultrasound 60 s, 0.2 s on, 2 s off, amplitude 22%) in lysis buffer (7M urea, 2M Thiourea, 4% CHAPS, and 50 ml protease inhibitor cocktail) and quantified via Bradford assay. 200 μg protein was taken from each sample and reduced and alkylated with 1 L of 25 mM DTT at 60 °C for 1 hour. Then, 0.5 L of 50 mM iodoacetamide was added and incubated for another 10 min at room temperature. Subsequently, these protein samples were centrifugated with 12,000 g for 20 min in 10K ultrafilter device (Milipore), then washed three times with 100 μl dissolution buffer from the iTRAQ Regents 8-plex kit (AB Sciex, Framingham,Ma,8390812) and centrifugated for 20 min at 12,000 g. Finally, 50 μl□4 μg□trypsin (Promega) was added into the ultrafilter device and incubated overnight at 37 °C to digest proteins.

On the following day, the digested peptide solution was collected by centrifugation at 12,000 g for 20 min and peptides were labeled with iTRAQ reagents according to the manufacturer's instructions. For each 100 μg of protein, one unit of labeling reagent was used, and labeling was performed for two hours and was stopped by adding 100 μl of water. The following labels were used for samples: 113, 115, and 116 were used to label East Asian strains GZ1, GZ2, and GZ3, while 118, 119, and 121 were used to label Western strains NCTC11635, GZ5, and 26695. iTRAQ-labeled samples were mixed and dried by vacuum freeze centrifugation. The dried samples were frozen until further use.

Reversed-phase chromatography pre-separation at high pH

The iTRAQ-labeled mixed sample was dissolved in 100 μl mobile phase A solvent (98 % dd H_2O, 2 % acetonitrile, pH 10) and loaded onto a Durashell-C18 column (4.6 mm× 250 mm, 5 μm, 100 Å, Agela). Peptides were eluted at 0.7 ml /min in a 306 min gradient using mobile phase A and B solvent (98% acetonitrile. 2 % dd H_2O, pH 10). The gradient was run as follows: 5% B over 0 min; 8% B over 5 min; 18% B over 35 min; 32 B % over 62 min; 95% B over 64 min; 95% B over 68 min; and 5% B over 72 min.

LC-MS/MS Analysis

Thermo Scientific EASY-nLC 1000 System (Nano HPLC) and Thermo Q-Exactive mass spectrometry system were employed to LC-MS/MS Analysis. In brief, the peptide components separated by the above reversed phase chromatography were re-dissolved with 20 μl 2% methanol and 0.1% formic acid and centrifuged for 10 min at 12,000 g. Then, 10 μl of supernatant was loaded onto an EASY-Spray column (LC column) and eluted with A solvent (100 % ddH2O, 0.1% formic acid) and B solvent (100 % acetonitrile, 0.1% formic acid). A gradient elution was used as follows: 5% B over 0 min twice; 8% B over 13 min; 30 % B over 90 min; 50 % B over 100 min; 95% B over 105; 95 % B over 115 min; 5% B over 116 min; and 5% B over 126 min. The flow rate of loading pump was 350 nl/min and the separation flow rate was 300nl/min.

The mass spectrometry parameters were set as follows. Ion source parameters were: Spray voltage: 2.3 kv; Capillary Temperature: 320 °C; Ion Source: EASY-Spray source; and DP: 100. Full MS parameters were: Resolution: 70000FWHM; Full Scan AGC target: 3e6; Full Scan Max.IT: 20 ms; and Scan range: 300-1800m/z. dd-MS2 parameters were: Resolution: 17500 FWHM; AGC target: 1e5; Maximum IT: 120ms; Intensity threshold: 8.30E+03; Fragmentation Methods: HCD; NCE: 32 %; and Top N: 20.

Database searching

Tandem mass spectra were extracted by ProteoWizard version 3.0.8789. All MS/MS samples were analyzed using Mascot (Matrix Science, London, UK; version 2.6.0). Mascot was set up to search the UniProt_*Helicobacter pylori* database (https://www.uniprot.org). Mascot was searched with a fragment ion mass tolerance of 0.020 Da and a parent ion tolerance of 10.0 PPM. Carbamidomethyl of cysteine and iTRAQ8plex of lysine and the N-terminus were specified in Mascot as fixed modifications. Oxidation of methionine, acetyl of the N-terminus and iTRAQ8plex of tyrosine were specified in Mascot as variable modifications.

Quantitative data analysis

Scaffold Q+ (version Scaffold_4.6.2, Proteome Software Inc., Portland, OR) was used to quantitate label-based quantitation peptide and protein identifications. Referring to the method reported in references [35,36], the probability computed by Protein Prophet algorithm was introduced to protein identification by MS/MS. Peptide identifications were accepted if they could be established at greater than 91% probability to achieve an FDR (false discovery rate) less than 1.0% by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at greater than 85% probability to achieve an FDR less than 10% and contained at least 1 identified peptide [37]. Protein probabilities were assigned by the Protein Prophet algorithm. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Normalization was performed iteratively (across samples and spectra) on intensities. Medians were used for averaging. Spectra data were log-transformed, pruned of those matched to multiple proteins, and weighted by an adaptive intensity weighting algorithm. Of 21287 spectra in the experiment at the given thresholds, 7668 (36%) were included in quantitation.

Identification of differentially expressed proteins

Six *H. pylori* strains were grouped into East Asian (3 strains) and Western groups (3 strains). The normalized intensity of each protein in the two groups was acquired from the above-described quantitative data analysis. The average of the normalized intensity of a single protein from the East Asian group or Western group was calculated, respectively, and the fold change (FC) of the single protein was defined as the ratio of Western group to East Asian group. The differentially expressed proteins of the Western group versus East Asian group were identified according to the criteria of Bonferroni-corrected $P \leq 0.01$, and fold change ≥ 1.2 (up-regulation) or ≤ 0.8 (down-regulation) [38,39].

Gene ontology (GO), KEGG pathway enrichment, and interaction network analysis

Gene ontology for differentially expressed proteins was conducted using UniProt database and DAVID 6.8 online analysis tool (https://david.ncifcrf.gov/) and visualized with GOplot R package. KEGG pathway enrichment was carried out via the KEGG pathway database and KOBAS 3.0 online tool (http://kobas.cbi.pku.edu.cn/); protein-protein interaction network (PPI) analysis was constructed using STRING 10.0 (https://string-db.org). Pathways and networks were visualized with Cytoscape software (version 3.7.1). GO items/pathways with FDR-corrected *P* value < 0.05 were considered significantly different [40].

RT-qPCR

mRNA levels of five differential proteins were determined by RT-qPCR. *16S rDNA* gene of *H. pylori* was used to normalize the expression level of the target genes. The prime sequences used in this study were list in Table 3. Total RNA of 10 East Asian strains and 10 Western strains were abstracted and transcribed into cDNA using the PrimeScript RT Reagent Kit with Gdna Eraser (TaKaRa, Dalian, China). qPCR was performed using SYBR Green I real-time PCR method (TaKaRa, Dalian, China) with twostep reactions according to the manufacturer's recommended protocol. The efficiency of the target amplification and the reference amplification were detected. $2^{-\Delta\Delta Ct}$ method was used to calculate the relative expression level of the target genes with East Asian group set as 1. The RT-qPCR analysis was performed with 10 biological repeats, and each sample had three technological repeats. The data were presented as the mean ± SD, and two-sided Student's t-test was used to perform statistical analysis using Graphpad Prism 8 program.

Abbreviations

CagA: Cytotoxin-associated gene A protein; iTRAQ: Isobaric tags for relative and absolute quantitation; UreA: Urease subunit alpha; UreH: Urease accessory protein H; FlaA: Flagellin A; FlgE: Flagellar hook protein E; FlhA: Flagellar biosynthesis protein A; FtsZ: Cell division protein Z; FtsI: Cell division protein I; GroEL: Chaperonin; CM: CagA multimerization; EPIYA: Glu-Pro-Ile-Tyr-Ala; ASR: Age-standardized rate; PS: Phosphatidylserine; NBS: N–terminal binding sequences; CBS: C-terminal binding sequences; PAR1b: Partitional defective 1 kinase b; PCA: Principal Component Analysis; FDR: False discovery rate; PPI: SD: Standard deviation; CV: Coefficient of variation.

Declarations

Acknowledgements

We would like to thank Dr. Fred Bogott at Austin Medical Center, Mayo Clinic in Austin, Minnesota, USA, for his excellent English editing of this manuscript.

Authors' contributions

ZJJ contributed to study concept and design, analysis, review and drafting of the manuscript. ZY and XY were responsible for the data collection of the manuscript. XL, QXY and ZQF were responsible for the data collection, analysis, and interpretation of the manuscript. WQR, WWL and ZQF helped with data collection, and review of the manuscript. LAZ and ZL contributed to study concept, critical review and revision of the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.31760328), the Guizhou Provincial Natural Science Foundation (Qiankehepintairencai [2017]5652 and Qiankehejichu [2016] 1416), the Science Foundation of Department of Education of Guizhou Province (QianJiaoHeKY[2018]020) and Guiyang Natural Science Foundation (Zhukehetong [2017]30-4)

Availability of data and materials

All sequencing data in this study have been uploaded to the NCBI database. The accession nomber is KR154731-KR154758, GQ161098, GQ161099, KR154758, and KR154757.

The mass spectrometry proteomics data in our manuscript submitted to “BMC genomics” have been deposited to the ProteomeXchange Consortium with the dataset identifier PXD020811 via the iProX partner repository. This dataset is available in https://www. iprox. org /page/SSV024.html;url=1597422004885nRzq (Password: N9rS).

Ethics approval and consent to participate

The study involving human participants were reviewed and approved by the Ethics Committee of Guiyang Medical University and all subjects provided their written informed consent in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Park JY, Forman D, Waskito LA, Yamaoka Y, Crabtree JE. Epidemiology of Helicobacter pylori and CagA-Positive Infections and Global Variations in Gastric Cancer. Toxins (Basel). 2018; 10: 1-20.
2. Zheng Q, Chen XY, Shi Y, Xiao SD. Development of gastric adenocarcinoma in 4Mongolian gerbils after long-term infection with Helicobacter pylori. J Gastroenterol Hepatol. 2004; 19:1192-8.

3. Brenner H, Arndt V, Stegmaier C, Ziegler H, Rothenbacher D. Is Helicobacter pylori infection a necessary condition for noncardia gastric cancer? Am J Epidemiol. 2004; 159: 252-8.
4. Plummer M, de Martel C, Vignat J, Ferlay J, Bray F, Franceschi S. Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob Health. 2016; 4: e609-16.
5. https://gco.iarc.fr/.
6. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68:394-424.
7. Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, Tanyingoh D, et al. Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. Gastroenterology. 2017;153: 420-9.
8. Peleteiro B, Bastos A, Ferro A, Lunet N. Prevalence of Helicobacter pylori infection worldwide: a systematic review of studies with national coverage. Dig Dis Sci. 2014; 59:1698-1709.
9. Abdullah M, Greenfield LK, Bronte-Tinkew D, Capurro MI, Rizzuti D, Jones NL. VacA promotes CagA accumulation in gastric epithelial cells during Helicobacter pylori infection. Sci Rep. 2019; 9: 38.
10. Backert S, Tegtmeyer N. Type IV Secretion and Signal Transduction of Helicobacter pylori CagA through Interactions with Host Cell Receptors. Toxins (Basel). 2017; 9:115.
11. Knorr J, Ricci V, Hatakeyama ., Backert S. Classification of Helicobacter pylori Virulence Factors: Is CagA a Toxin or Not? Mol Cell Proteomics. 2019;27:731-8.
12. Shiota S, Matsunari O, Watada M, Yamaoka Y. Serum Helicobacter pylori CagA antibody as a biomarker for gastric cancer in east-Asian countries. Future Microbiol. 2010; 5:1885-93.
13. Hatakeyama M. Helicobacter pylori CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. Cell Host Microbe. 2014;15:306-16.
14. Nagase L, Hayashi T, Senda T, Hatakeyama M. Dramatic increase in SHP2 binding activity of Helicobacter pylori Western CagA by EPIYA-C duplication: its implications in gastric carcinogenesis. Sci Rep. 2015; 5: 15749.
15. Li Q, Liu J, Gong Y, Yuan Y. Association of CagA EPIYA-D or EPIYA-C phosphorylation sites with peptic ulcer and gastric cancer risks: A meta-analysis. Medicine. 2017;96:e6620.
16. Hatakeyama, M. Oncogenic mechanisms of the Helicobacter pylori CagA protein. Nat Rev Cancer. 2004;4:688-94.
17. Nejati S, Karkhah A, Darvish H, Validi M, Ebrahimpour S, Nouri HR. Influence of Helicobacter pylori virulence factors CagA and VacA on pathogenesis of gastrointestinal disorders. Microb Pathog. 2018; 117: 43-8.
18. Ding SZ, Goldberg JB, Hatakeyama M. Helicobacter pylori infection, oncogenic pathways and epigenetic mechanisms in gastric carcinogenesis. Future Oncol. 2010;6:851-62.
19. Chen SY, Zhang RG, Duan GC. Pathogenic mechanisms of the oncoprotein CagA in H. pylori-induced gastric cancer (Review). Oncol Rep. 2016; 36: 3087-94.

20. Noto JM, Rose KL, Hachey AJ, Delgado AG, Romero-Gallo J, Wroblewski LE, et al. Carcinogenic Helicobacter pylori Strains Selectively Dysregulate the In Vivo Gastric Proteome, Which May Be Associated With Stomach Cancer Progression. Mol Cell Proteomics. 2019;18: 352-71.
21. Müller SA, Pernitzsch SR, Haange SB, Uetz P, von Bergen M, Sharma CM, et al. Stable isotope labeling by amino acids in cell culture based proteomics reveals differences in protein abundances between spiral and coccoid forms of the gastric pathogen Helicobacter pylori. J Proteomics. 2015; 126: 34-45.
22. Martin B, Brenneman R, Becker KG, Gucek M, Cole RN, Maudsley S. iTRAQ analysis of complex proteome alterations in 3xTgAD Alzheimer's mice: understanding the interface between physiology and disease. PLoS One. 2008; 3: e2750.
23. Ohnishi N, Yuasa H, Tanaka S, Sawa H, Miura M, Matsui A, et al. Transgenic expression of Helicobacter pylori CagA induces gastrointestinal and hematopoietic neoplasms in mouse. Proc. Natl. Acad. Sci. USA. 2008; 105: 1003-8.
24. Jones TA, Hernandez DZ, Wong ZC, Wandler AM, Guillemin K. The bacterial virulence factor CagA induces microbial dysbiosis that contributes to excessive epithelial cell proliferation in the Drosophila gut. PLoS Pathog. 2017; 13, e1006631.
25. Pormohammad A, Ghotaslou R, Leylabadlo HE, Nasiri, MJ, Dabiri, H, Hashemi A. Risk of gastric cancer in association with Helicobacter pylori different virulence factors: A systematic review and meta-analysis. Microb Pathog. 2018; 118: 214-9.
26. Hatakeyama M. Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer. Proc Jpn Acad Ser B Phys Biol Sci. 2017; 93: 196-219.
27. Nishikawa H, Hatakeyama M. Sequence Polymorphism and Intrinsic Structural Disorder as Related to Pathobiological Performance of the Helicobacter pylori CagA Oncoprotein. Toxins (Basel). 2017; 9: 136.
28. Zhang XS, Tegtmeyer ., Traube L, Jindal S, Perez-Perez G, Sticht H, et al. A specific A/T polymorphism in Western tyrosine phosphorylation B-motifs regulates Helicobacter pylori CagA epithelial cell interactions. PLoS Pathog. 2015; 11: e1004621.
29. Batista SA, Rocha GA, Rocha AM, Saraiva IE, Cabral MM, Oliveira RC, et al. Higher number of Helicobacter pylori CagA EPIYA C phosphorylation sites increases the risk of gastric cancer, but not duodenal ulcer. BMC Microbiol. 2011; 11: 61
30. Gupta N, Maurya S, Verma H, Verma VK. Unraveling the factors and mechanism involved in persistence: Host-pathogen interactions in Helicobacter pylori. J Cell Biochem. 2019;120:18572-87.
31. Lowenthal AC, Hill M, Sycuro LK, Mehmood K, Salama NR, Ottemann KM. Functional analysis of the Helicobacter pylori flagellar switch proteins. J Bacteriol. 2009; 191: 7147-56.
32. Gu H. Role of Flagella in the Pathogenesis of Helicobacter pylori. Curr Microbiol. 2017; 74: 863-9.
33. Fong YH, Wong HC, Yuen MH, Lau PH, Chen YW, Wong KB. Structure of UreG/UreF/UreH complex reveals how urease accessory proteins facilitate maturation of Helicobacter pylori urease. PLoS Biol. 2013; 11: e1001678.

34. Huang JY, Sweeney EG, Sigal M, Zhang HC, Remington SJ, Cantrell MA, et al. Chemodetection and Destruction of Host Urea Allows Helicobacter pylori to Locate the Epithelium. Cell Host Microbe. 2015; 18: 147-56.
35. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75:4646-58.
36. Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, et al. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res. 2008;7:225-33.
37. Higdon R, Kolker E. A predictive model for identifying proteins by a single peptide match. Bioinformatics. 2007; 233: 277–80.
38. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009; 27:1135-7.
39. Gao Y, Long R, Kang J, Wang Z, Zhan T, Sun H, Li X, Yang Q. Comparative Proteomic Analysis Reveals That Antioxidant System and Soluble Sugar Metabolism Contribute to Salt Tolerance in Alfalfa (Medicago sativa L) Leaves. J Proteome Res. 2019; 18:191-203.
40. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Statist Soc. B. 1995; 57: 289-300.

Tables

Table 1 Characteristics of EPIYA motifs in CagA from clinical isolates

Strains	GenBank accession no	Diseases of strain origin	EPIYA motifs	Classification	Variation	
					site	type
GZ1	KR154731	Chronic gastritis	ABD	East Asia	No	
GZ2	KR154732	Gastric ulcer	ABD	East Asia	No	
GZ3	KR154733	Atrophic gastritis	ABD	East Asia	No	
GZ4	KR154734	Atrophic gastritis	ABD	East Asia	No	
GZ7	KR154737	Gastric cancer	ABD	East Asia	No	
GZ8	KR154738	Gastric ulcer	AB**D**[#]	East Asia	EPIY**A**-D	A→V
GZ11[a]	KR154741	Chronic gastritis	ABD	East Asia	No	
GZ12	KR154742	Chronic gastritis	ABD	East Asia	No	
GZ13	KR154743	Gastric ulcer	ABD	East Asia	No	
GZ16	KR154745	Gastric ulcer	ABD	East Asia	No	
GZ17[b]	KR154746	Gastritis	ABD	East Asia	No	
GZ18[b]	KR154747	Gastritis	ABD	East Asia	No	
GZ19	KR154748	Gastritis	A**B**[#]D	East Asia	E**P**IYA-B	P→S
GZ20	KR154749	Gastric ulcer	ABD	East Asia	No	
GZ21	KR154750	Chronic gastritis	ABD	East Asia	No	
GZ23	KR154752	Gastritis	ABD	East Asia	No	
GZ24	KR154753	Gastritis	ABD	East Asia	No	
GZ25	KR154754	Gastrorrhagia	ABD	East Asia	No	
GZ26	KR154755	Gastric ulcer	AB**D**[#]	East Asia	EPIY**A**-D	A→V
GZ27	KR154756	Atrophic gastritis	BD	East Asia	EPIYA-A	Deletion
GZ15[c]	-	Gastric ulcer	-	-	EPIYA	Deletion
GZ5	KR154735	Chronic gastritis	A**B**[#]C	Western	EPIYA-B	A→T

GZ6	KR154736	Gastric cancer	A**B**[#]C	Western	EPIY**A**-B	A→T
GZ9	KR154739	Gastric ulcer	AB	Western	EPIYA-C	Deletion
GZ10[a]	KR154740	Chronic gastritis	A **B**[#]C	Western	EPIY**A**-B	A→T
GZ14	KR154744	Gastritis	AB	Western	EPIYA-C	Deletion
GZ22	KR154751	Gastric ulcer	AB	Western	EPIYA-C	Deletion
SS1[d]	KR154757	-	AB	Western	EPIYA-C	Deletion
NCTC11637[d]	GQ161098	-	AB	Western	EPIYA-C	Deletion
NCTC11639[d]	GQ161099	-	A**B**[#]C	Western	EPIY**A**-B	A→T
26695[d]	KR154758	-	A**B**[#]C	Western	EPIY**A**-B	A→T

ABD/C indicates EPIYA-A, -B, and -D/C sites. Grey shadow represents the amino acid variation. a and b represent strains from the same patient. [c]: A clinical isolate with deletion of C-terminal region of CagA was not classified into any group and its sequence was not submitted to the GenBank database. [d]: Gifts from *H. pylori* Strain Management and Preservation Center, China. [#]: Variation site.

Table 2 Sequence variations in EPIYA motifs of CagA of 150 *H. pylori* strains from GenBank

EPIYA motifs	Strains (%)			Variation
	Total (n=150)	East Asian (n=80)	Western (n=70)	
EPIYA-A	2 (1.3)	1 (1.3)	1 (1.4)	Deletion
E**P**IYA-B	5 (3.3)	3 (3.8)	2 (2.9)	P→S
EPIY**A**-B[a]	29 (19.3)	4 (5)	25 (35.7)	A→T
EP**I**YA-B[b]	1 (0.6)	0	1 (1.4)	I→V
EPIY**A**-B[c]	1 (0.6)	1 (1.3)	0	A→S
EPIYA-C	31 (20.7)	0	31 (44.3)	Duplication
EPIYA-C	6 (4)	1 (1.3)	5 (7.1)	Deletion
EPIY**A**-D	2 (1.3)	2 (2.5)	0	A→V
Sum	73 (48.7)	11(13.8)	62 (88.6)	

a: *H. pylori* strains Shi257 (AB587256.1) and Shi417 (AB587258.1) contain the A→T and P→S variations simultaneously. b: This strain contains the I→V and A→T variations simultaneously. c: This strain contains the P→S and A→S variations simultaneously.

Table 3 Primer sequences used in this study

Name	Primer (5'-3'）	GenBank accession ID	Size(bp)
ureA	Sense：GCCATGAAAACCACGCTCTT	WP_000779228.1	180
	Antisense：TTAGACATTGCGAGCGGGAC		
ureH	Sense：CAACTCATTGCACGCCACTC	WP_001099421.1	127
	Antisense：GCCTTTCCCGTTAATCCCCT		
flaE	Sense：TGGCAGAAACCACGGTGAAT	WP_000885474.1	130
	Antisense：GTTGTAACCCCAGCTCCCAA		
hpaA	Sense：GCCGCAATTCCACTCTTTCA	WP_041049496.1	213
	Antisense：AAGCGTTGCTCGTTTTACGC		
flgE	Sense：TGCGGTAGCGATGAGTTTGA	WP_033614689.1	131
	Antisense：GGATGCAAGCCCACCAAATC		
16S rDNA	Sense：CTCGGAATCACTGGGCGTAA	NR_074476.2	121
	Antisense：CCACCTACCTCTCCCACACT		

ureA: urease subunit alpha. *ureH*: urease accessory protein. *flaE*: flagellin. *hpaA*: hydroxyphenyl- acetate catabolism regulator. *flgE*: flagellar hook protein.

Figures

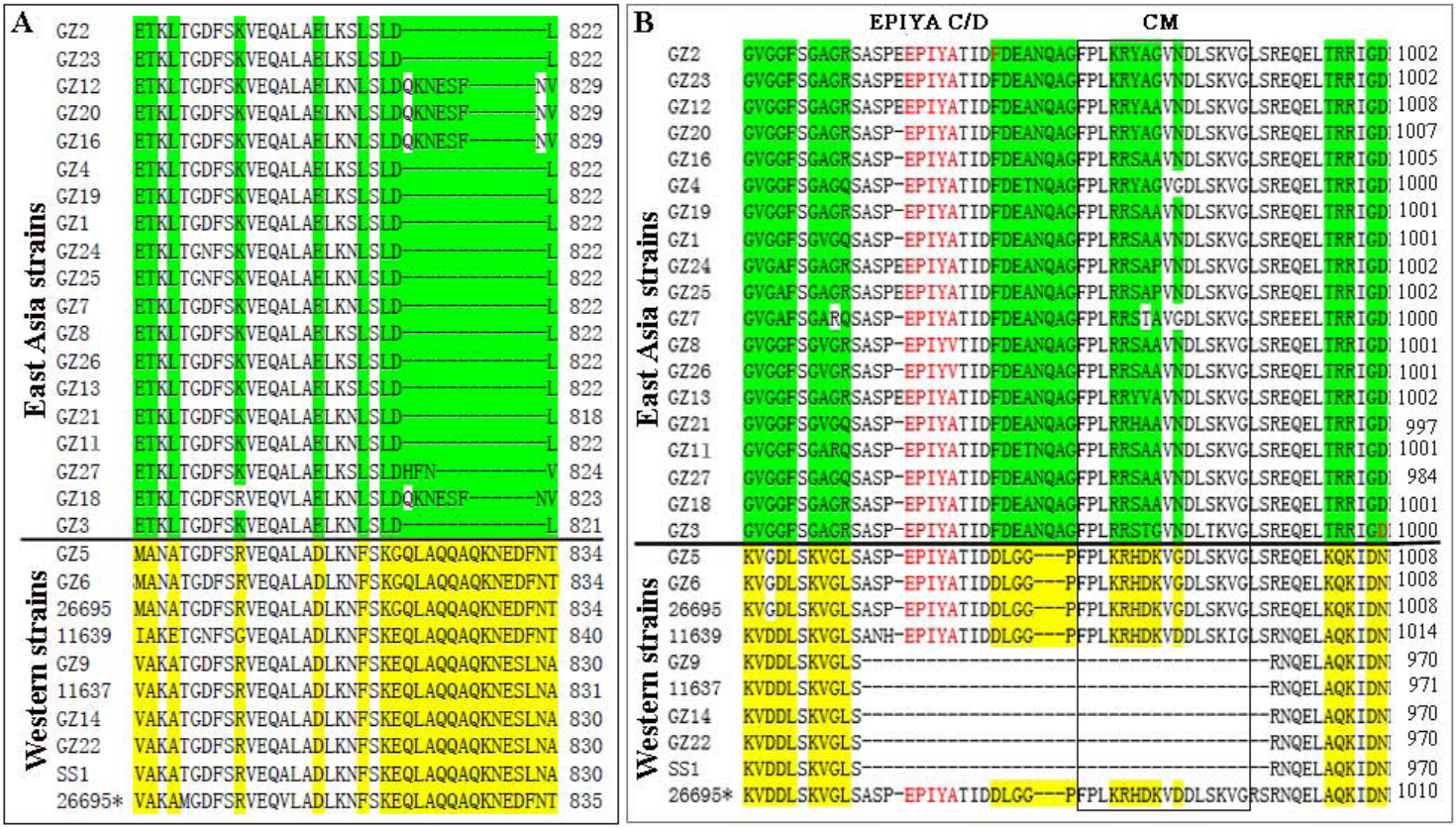


Figure 1

Amino acid sequence comparison of CagA at indicated regions between East Asian and Western strains. a, b The region at the 797th-835th (a) and 955th-1010th (b) amino acid sequences of CagA. Green and yellow shadow represents the differential amino acids in East Asian and Western CagA, respectively. Red letters represent EPIYA-C/D site. * indicates the CagA sequences of H. pylori strain 26695 from GenBank ID AAD07614.1 and as a reference sequence.

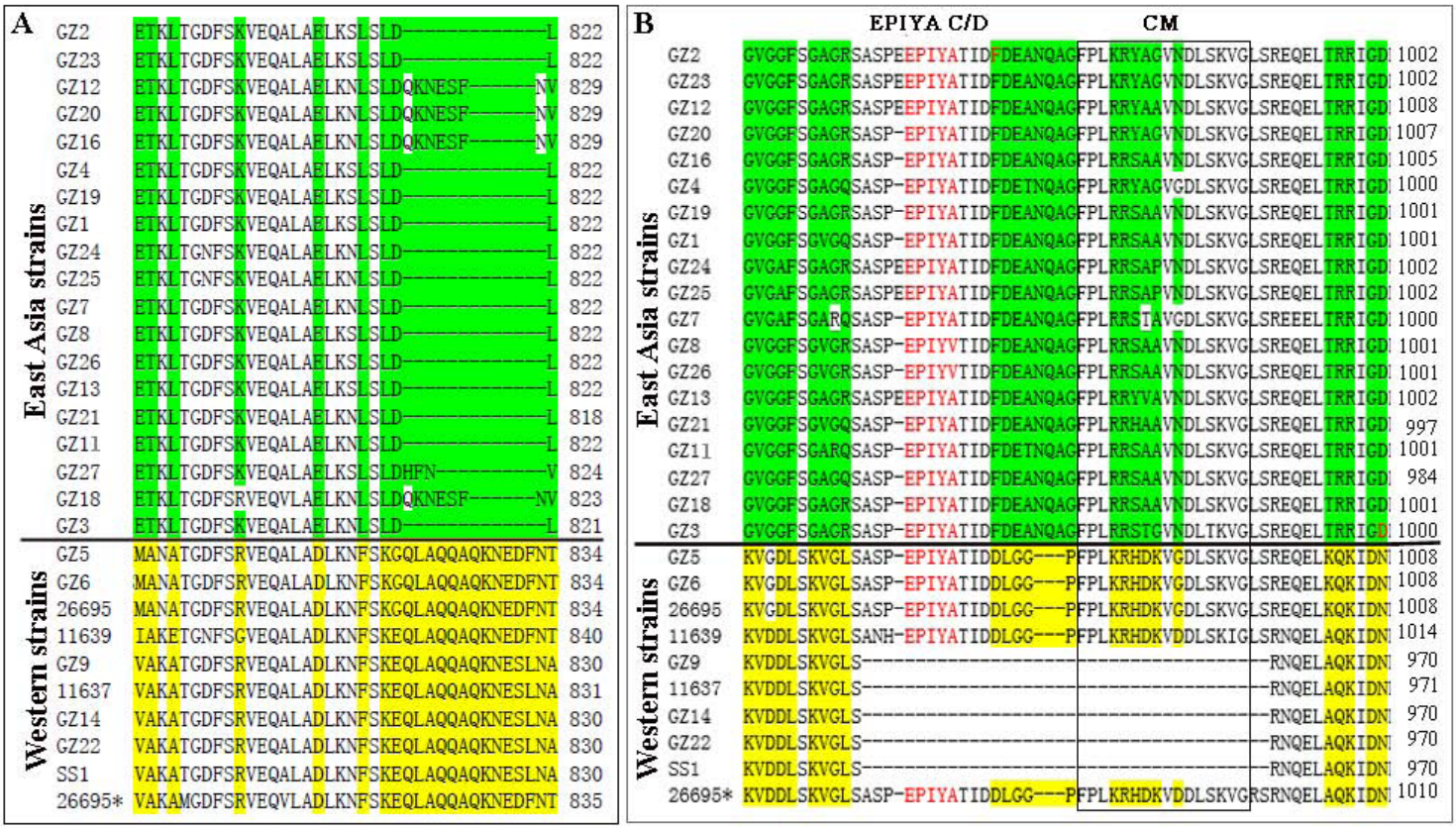


Figure 1

Amino acid sequence comparison of CagA at indicated regions between East Asian and Western strains. a, b The region at the 797th-835th (a) and 955th-1010th (b) amino acid sequences of CagA. Green and yellow shadow represents the differential amino acids in East Asian and Western CagA, respectively. Red letters represent EPIYA-C/D site. * indicates the CagA sequences of H. pylori strain 26695 from GenBank ID AAD07614.1 and as a reference sequence.

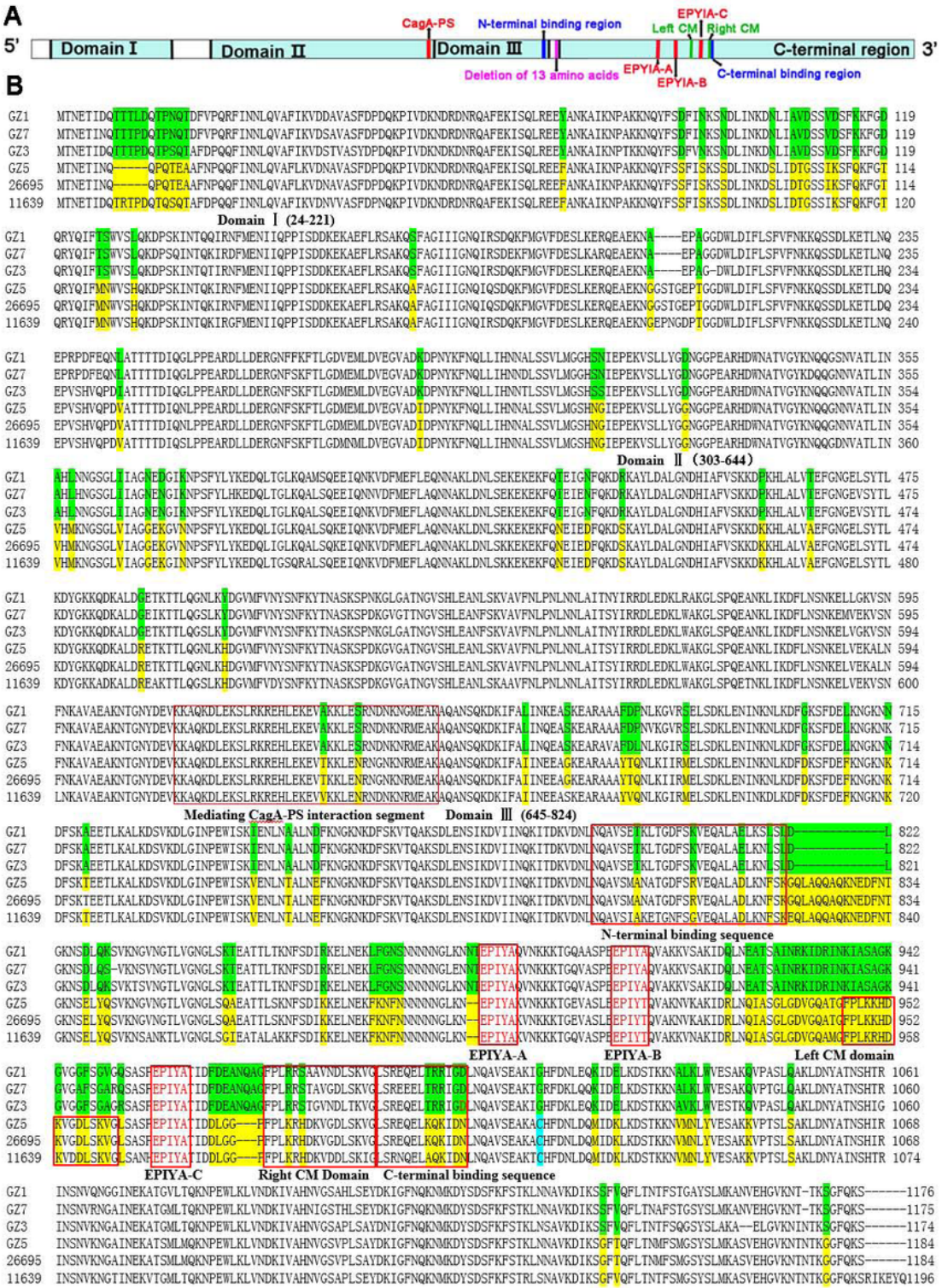


Figure 2

Sequence comparison of the functional domains of CagA between six representative East Asian and Western strains. a Linear schematic of CagA structure. CagA-PS represents their interaction segment. b Comparison of CagA sequences between East Asian and Western strains. Green and yellow shadow represents the differential amino acids of East Asian and Western CagA, respectively. Red box indicates the important domains of CagA. Blue shadow represents Cys. PS: phosphatidylserine.

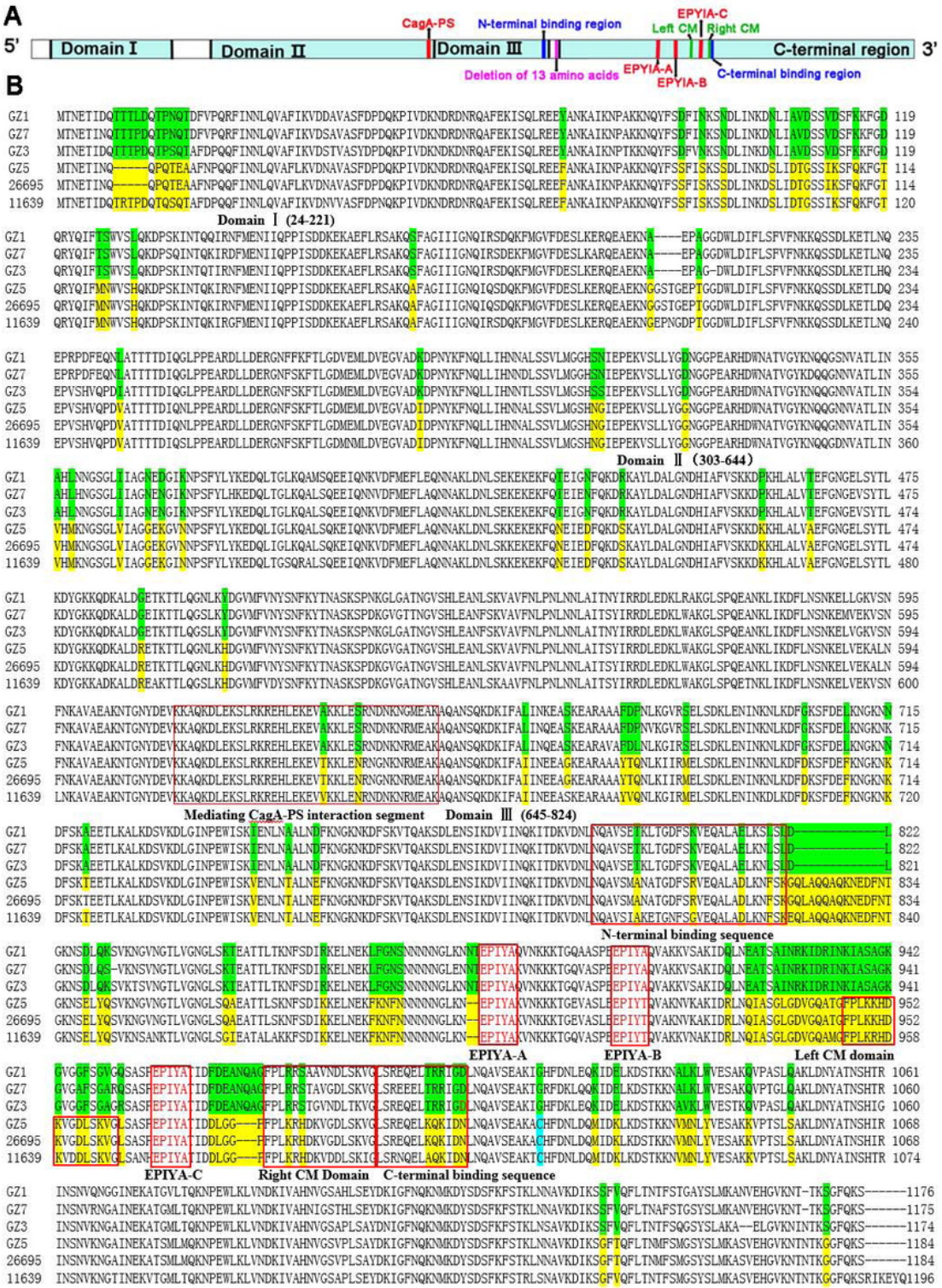


Figure 2

Sequence comparison of the functional domains of CagA between six representative East Asian and Western strains. a Linear schematic of CagA structure. CagA-PS represents their interaction segment. b Comparison of CagA sequences between East Asian and Western strains. Green and yellow shadow represents the differential amino acids of East Asian and Western CagA, respectively. Red box indicates the important domains of CagA. Blue shadow represents Cys. PS: phosphatidylserine.

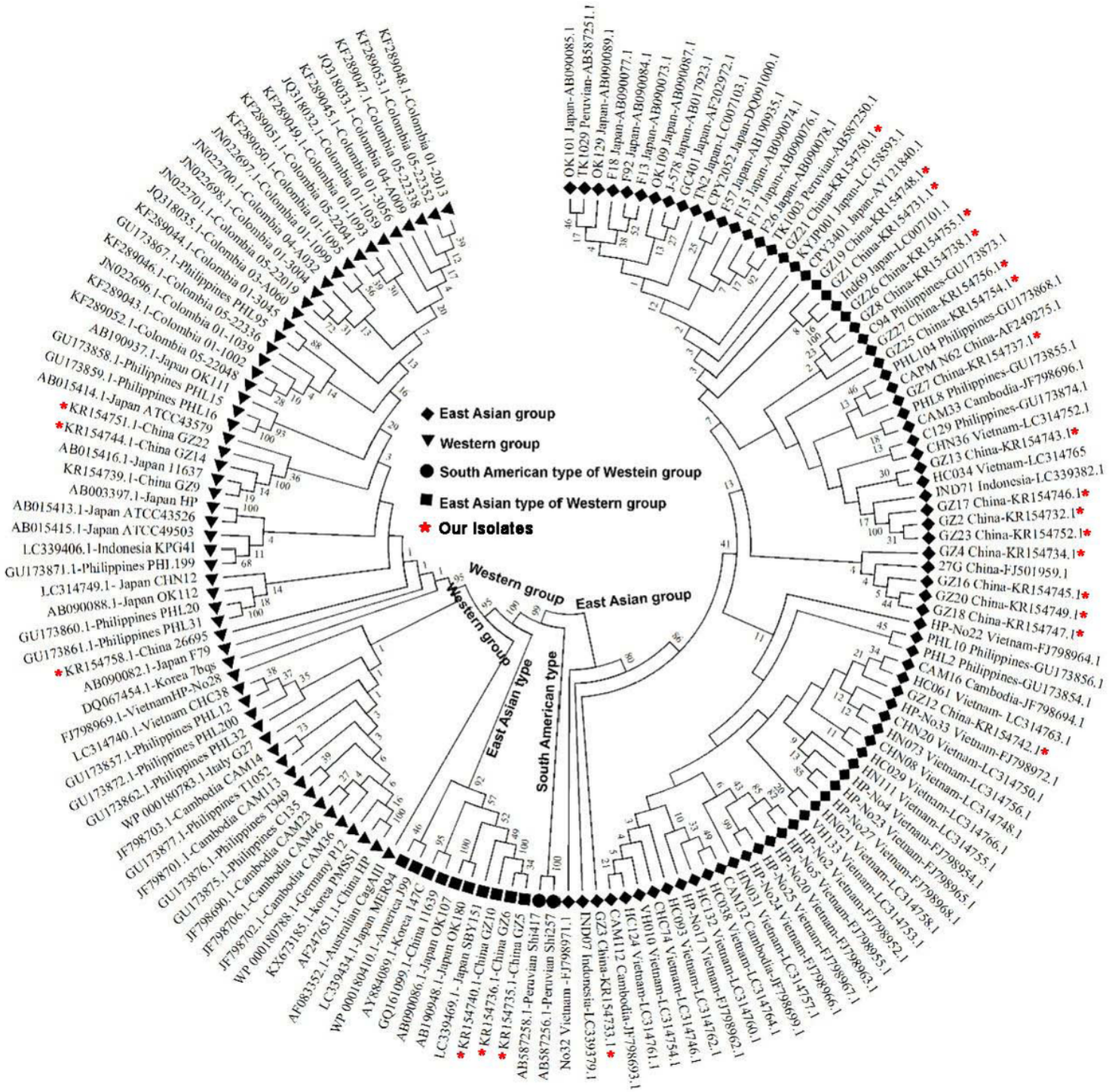


Figure 3

Phylogenic tree of 150 H. pylori stains based on CagA sequences. One hundred and fifty sequences of H. pylori-CagA were obtained from the Genbank database to construct the CagA-based molecular phylogenic tree.

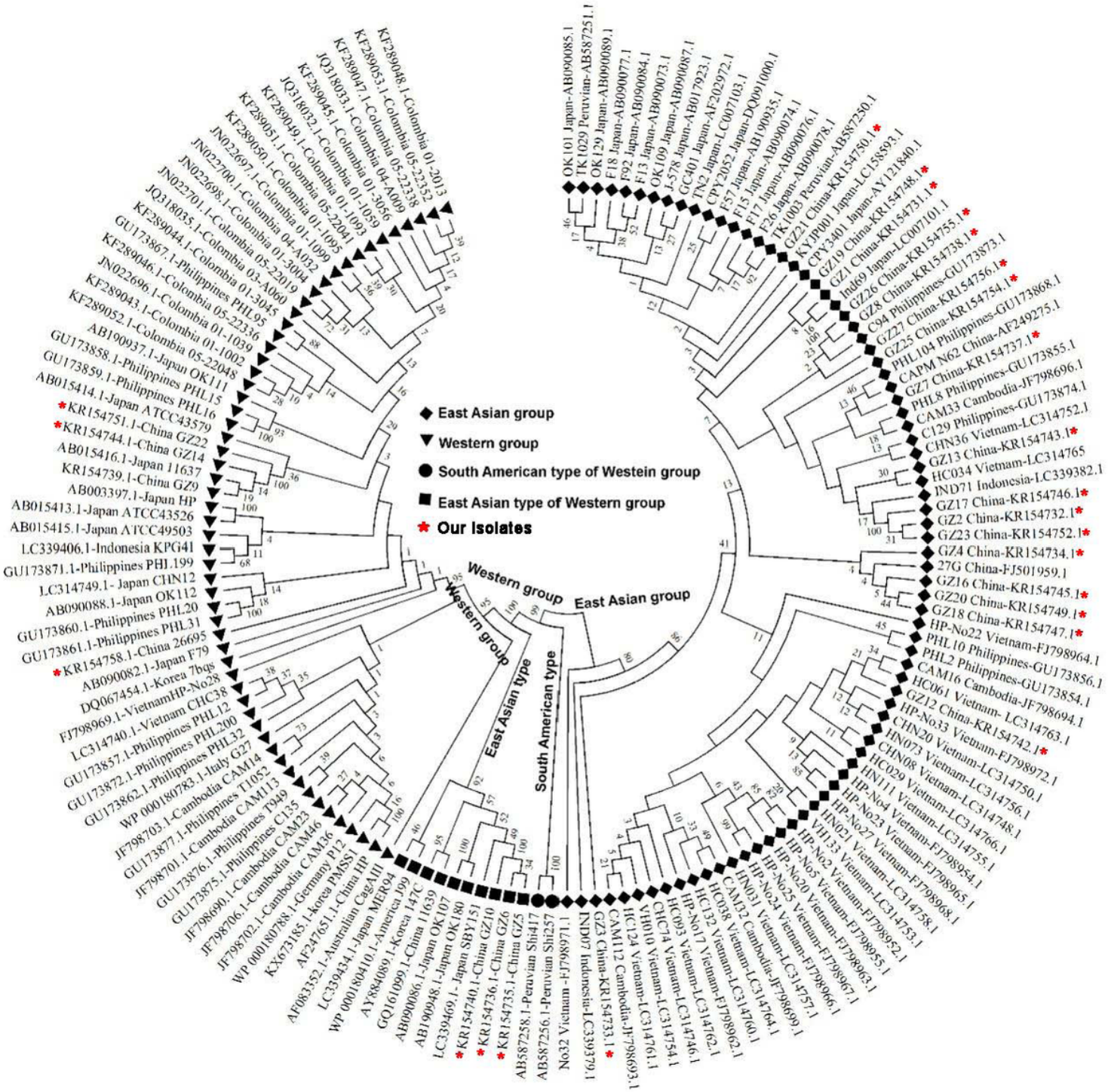


Figure 3

Phylogenic tree of 150 H. pylori stains based on CagA sequences. One hundred and fifty sequences of H. pylori-CagA were obtained from the Genbank database to construct the CagA-based molecular phylogenic tree.

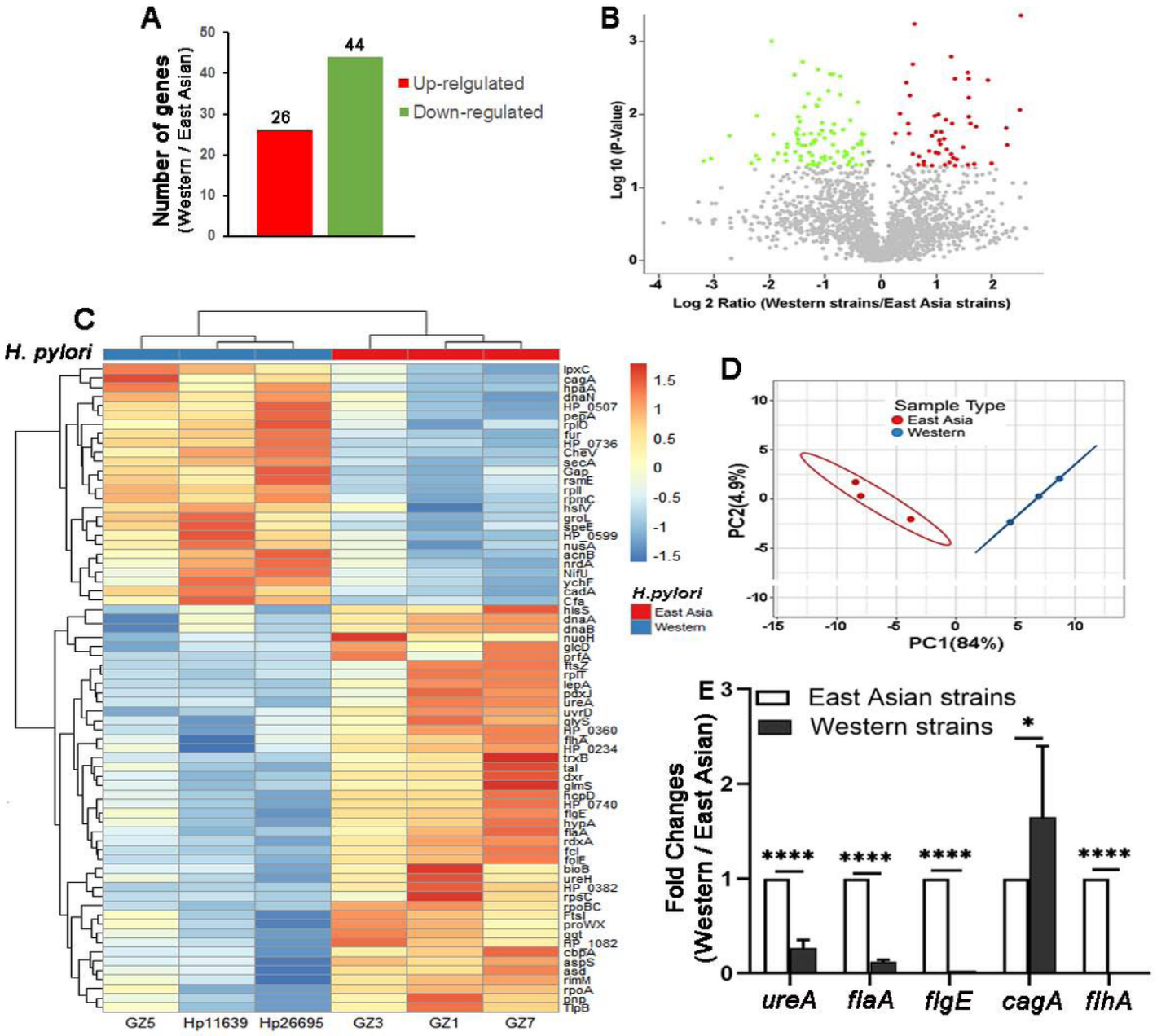


Figure 4

Differentially expressed proteins between East Asian and Western H. pylori strains. Three East Asian and three Western stains of H. pylori were selected to conduct iTRAQ-based quantitative proteomics analysis. a Differentially expressed proteins between Western and East Asian groups. b Volcano plot of differential proteins in the Western group compared to the East Asian group. Red and green plots represent up-regulation and down-regulation, respectively, in the Western group compared to the East Asian group. c Hierarchical clustering graph of differential proteins between the Western and East Asian group. Six H. pylori strains, including three East Asian strains (GZ3, GZ1 and GZ7) and three Western strains (GZ5, Hp11639 and Hp26695), were simultaneously collected for iTRAQ-based LC-MS/MS analysis. The proteomics data were derived from three independent biological experiments. Red color represents up-

regulation and blue color represents down-regulation. d Principal Component Analysis (PCA) of six samples. e mRNA levels of five differential proteins were determined by RT-qPCR. Total RNA of 10 East Asian strains and 10 Western strains were abstracted and transcribed into cDNA using the PrimeScript RT Reagent Kit with Gdna Eraser. qPCR was performed using SYBR Green I real-time PCR method with twostep reactions. 2−ΔΔCt method was used to calculate the relative expression level of the target genes with East Asian group set as 1. The RT-qPCR analysis was performed with 10 biological repeats, and each sample had three technological repeats. The data were presented as the mean ± SD, and two-sided Student's t-test was used to perform statistical analysis using Graphpad Prism 8 program. ureA: urease subunit alpha. ureH: urease accessory protein. flaE: flagellin. hpaA: hydroxyphenyl- acetate catabolism regulator. flgE: flagellar hook protein. * and **** represent p<0.01 and p<0.0001, respectively.

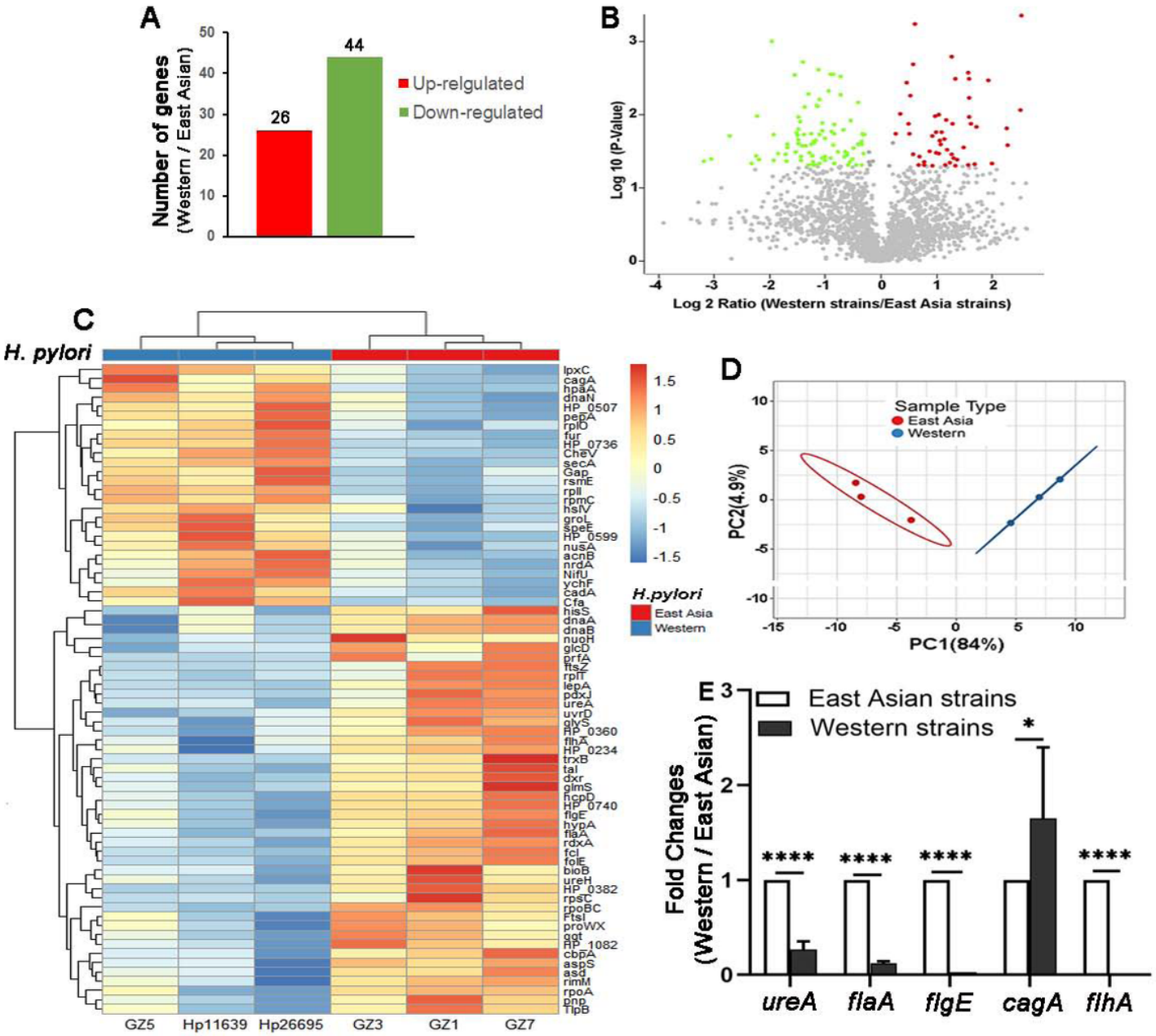
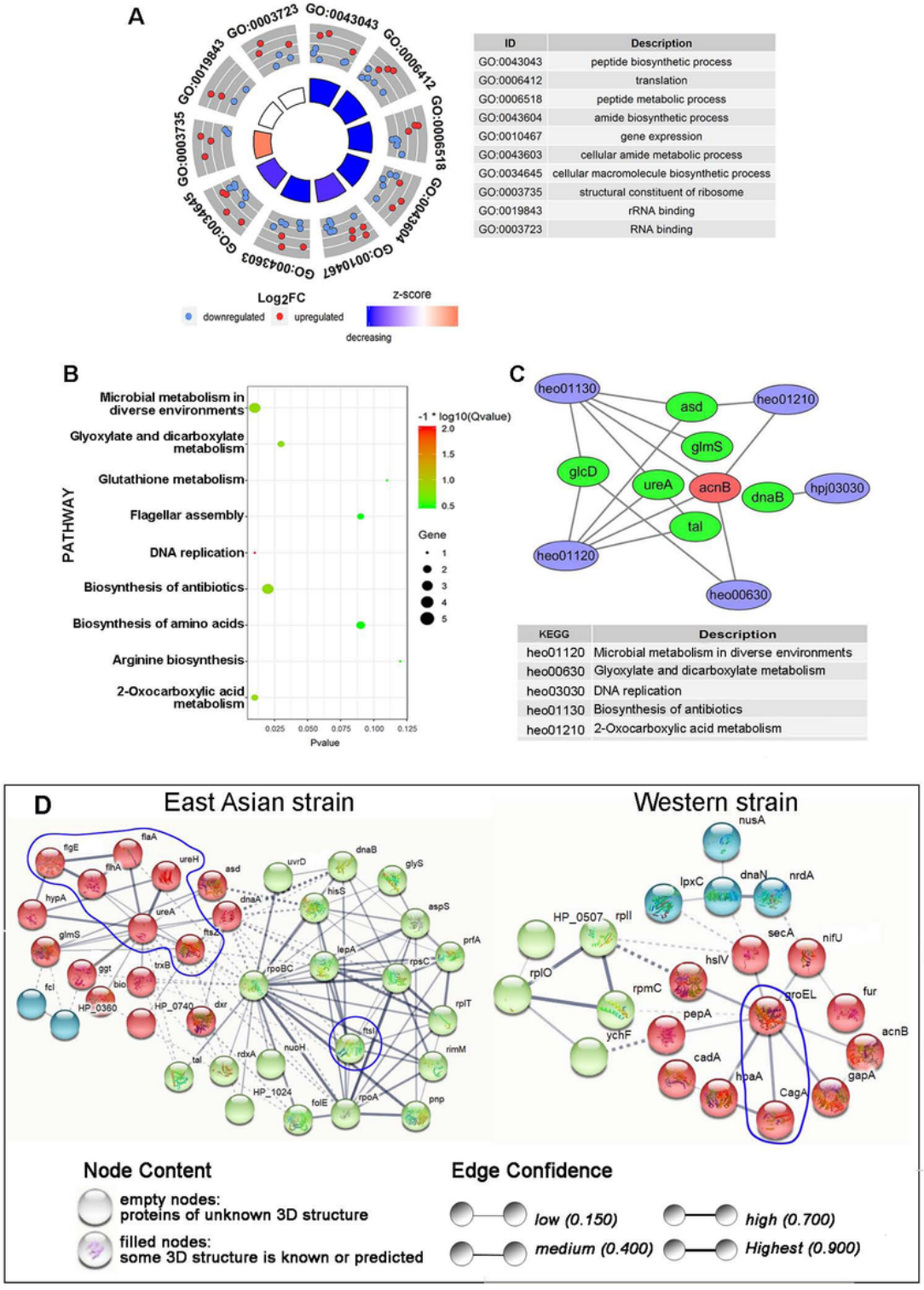


Figure 4

Differentially expressed proteins between East Asian and Western H. pylori strains. Three East Asian and three Western stains of H. pylori were selected to conduct iTRAQ-based quantitative proteomics analysis. a Differentially expressed proteins between Western and East Asian groups. b Volcano plot of differential proteins in the Western group compared to the East Asian group. Red and green plots represent up-regulation and down-regulation, respectively, in the Western group compared to the East Asian group. c Hierarchical clustering graph of differential proteins between the Western and East Asian group. Six H. pylori strains, including three East Asian strains (GZ3, GZ1 and GZ7) and three Western strains (GZ5, Hp11639 and Hp26695), were simultaneously collected for iTRAQ-based LC-MS/MS analysis. The proteomics data were derived from three independent biological experiments. Red color represents up-regulation and blue color represents down-regulation. d Principal Component Analysis (PCA) of six samples. e mRNA levels of five differential proteins were determined by RT-qPCR. Total RNA of 10 East Asian strains and 10 Western strains were abstracted and transcribed into cDNA using the PrimeScript RT Reagent Kit with Gdna Eraser. qPCR was performed using SYBR Green I real-time PCR method with twostep reactions. $2-\Delta\Delta Ct$ method was used to calculate the relative expression level of the target genes with East Asian group set as 1. The RT-qPCR analysis was performed with 10 biological repeats, and each sample had three technological repeats. The data were presented as the mean ± SD, and two-sided Student's t-test was used to perform statistical analysis using Graphpad Prism 8 program. ureA: urease subunit alpha. ureH: urease accessory protein. flaE: flagellin. hpaA: hydroxyphenyl- acetate catabolism regulator. flgE: flagellar hook protein. * and **** represent $p<0.01$ and $p<0.0001$, respectively.

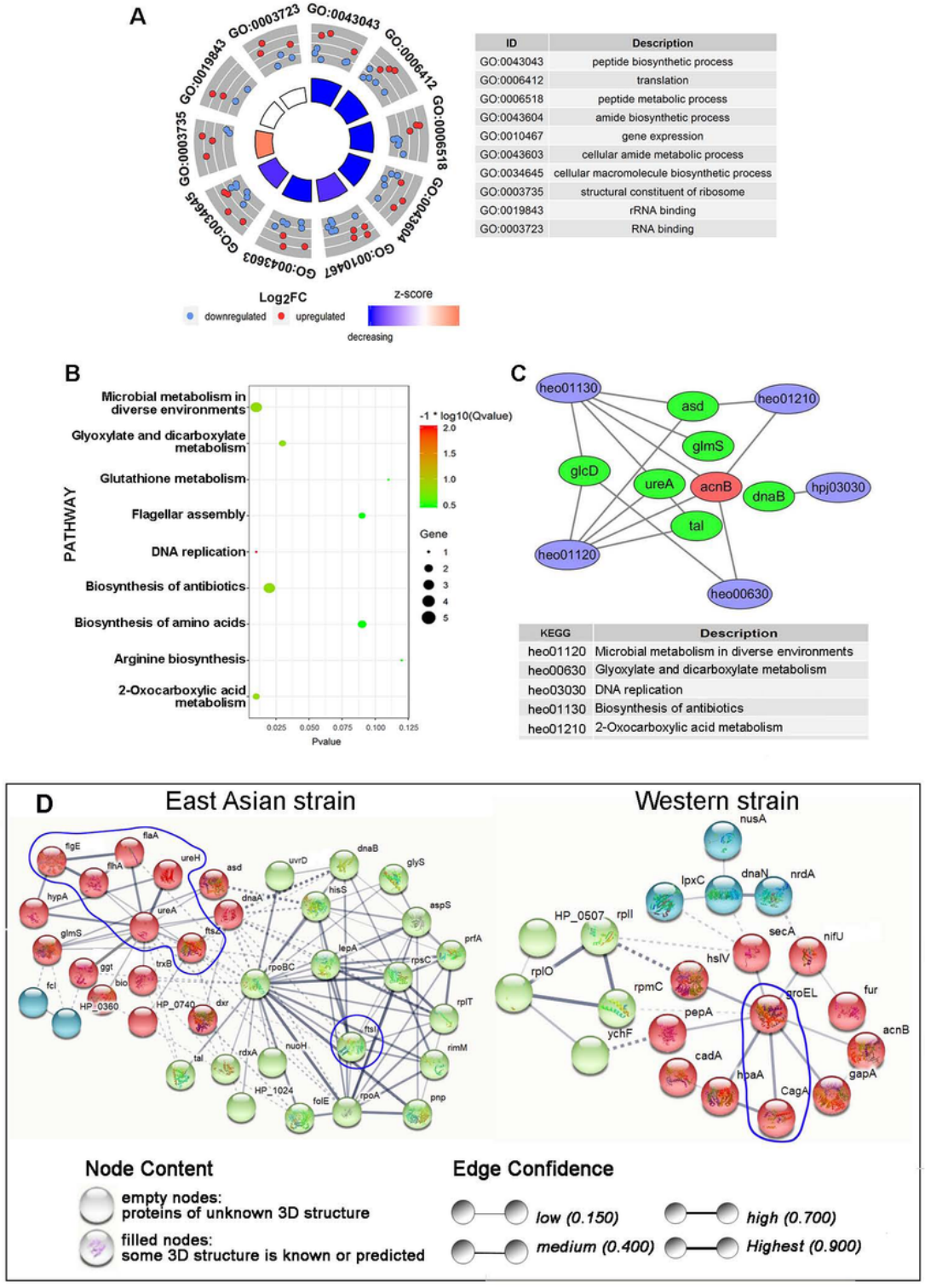


ID	Description
GO:0043043	peptide biosynthetic process
GO:0006412	translation
GO:0006518	peptide metabolic process
GO:0043604	amide biosynthetic process
GO:0010467	gene expression
GO:0043603	cellular amide metabolic process
GO:0034645	cellular macromolecule biosynthetic process
GO:0003735	structural constituent of ribosome
GO:0019843	rRNA binding
GO:0003723	RNA binding

KEGG	Description
heo01120	Microbial metabolism in diverse environments
heo00630	Glyoxylate and dicarboxylate metabolism
heo03030	DNA replication
heo01130	Biosynthesis of antibiotics
heo01210	2-Oxocarboxylic acid metabolism

Figure 5

Functional annotation and protein-protein interaction maps of 70 differentially expressed proteins. a GO annotation based on DAVID database. b, c KEGG pathway enrichment（b）and significant enrichment (c). d Protein-protein interactions (PPIs) of highly expressed proteins in East Asian and Western strains. Solid lines represent the establish interaction, and dashed lines indicate the predicted interaction. The line thickness and depth of line color (Edge Confidence) represent the levels of interaction.



ID	Description
GO:0043043	peptide biosynthetic process
GO:0006412	translation
GO:0006518	peptide metabolic process
GO:0043604	amide biosynthetic process
GO:0010467	gene expression
GO:0043603	cellular amide metabolic process
GO:0034645	cellular macromolecule biosynthetic process
GO:0003735	structural constituent of ribosome
GO:0019843	rRNA binding
GO:0003723	RNA binding

KEGG	Description
heo01120	Microbial metabolism in diverse environments
heo00630	Glyoxylate and dicarboxylate metabolism
heo03030	DNA replication
heo01130	Biosynthesis of antibiotics
heo01210	2-Oxocarboxylic acid metabolism

Figure 5

Functional annotation and protein-protein interaction maps of 70 differentially expressed proteins. a GO annotation based on DAVID database. b, c KEGG pathway enrichment（b）and significant enrichment (c). d Protein-protein interactions (PPIs) of highly expressed proteins in East Asian and Western strains. Solid lines represent the establish interaction, and dashed lines indicate the predicted interaction. The line thickness and depth of line color (Edge Confidence) represent the levels of interaction.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile.pdf
- Additionalfile.pdf