Diversity of 3' Variable Region of cagA Gene in Helicobacter Pylori Strains Isolated from Chinese Population

zhijing Xue
Chinese Center for Disease Control and Prevention https://orcid.org/0000-0002-9459-6663
Yuanhai You
Chinese Center for Disease Control and Prevention
Lihua He
Chinese Center for Disease Control and Prevention
Yanan Gong
Chinese Center for Disease Control and Prevention
Lu Sun
Chinese Center for Disease Control and Prevention
Xiurui Han
Chinese Center for Disease Control and Prevention
Ruyue Fan
Chinese Center for Disease Control and Prevention
Kangle Zhai
Chinese Center for Disease Control and Prevention
Yaming Yang
Chinese Center for Disease Control and Prevention
Maojun Zhang
Chinese Center for Disease Control and Prevention
Xiaomei Yan
Chinese Center for Disease Control and Prevention
Jianzhong Zhang (zhangjianzhong@icdc.cn)
Chinese Center for Disease Control and Prevention https://orcid.org/0000-0001-7056-8206

Research

Keywords: Helicobacter pylori, cagA, EPIYA, gastroduodenal disease, polymorphism

DOI: https://doi.org/10.21203/rs.3.rs-127429/v1

Abstract

Background: CagA is one of the most important virulence factors of *Helicobacter pylori* (*H. pylori*). There is a highly polymorphic Glu-Pro-Ile-Tyr-Ala (EPIYA) repeat region in the CagA 3' variable region. This repeat region is thought to play an important role in the pathogenesis of gastrointestinal diseases. The aim of this study was to investigate the diversity of CagA 3' variable region and the amino acid polymorphisms in the EPIYA segments, and their association with gastroduodenal diseases.

Methods: A total of 515 *H. pylori* isolates from patients in 14 different geographical regions of China were collected and the genomic DNA was extracted. The 3' variable region of the *cagA* was amplified by polymerase chain reaction (PCR) and then followed by DNA sequencing, and the amino acid sequences were analyzed with MEGA 7.0 software.

Results: A total of 503 (97.7%) *H. pylori* isolates were *cagA*-positive and 1,587 EPIYA motifs were obtained, including 12 types of EPIYA or EPIYA-like sequences. In addition to the four reported major segments, several rare segments (e.g., B', B" and D') were defined and 20 different sequence types (e.g., ABD, ABC) were found in our study. A total of 481 (95.6%) strains were East Asian type, most of them were ABD subtype (82.1%). Only 22 strains were Western type, including types AC, ABC, ABCC and ABCCCC. The CagA-ABD subtype had statistical difference in different geographic regions (P=0.006). There are seven amino acid polymorphisms in the sequences surrounding the EPIYA motifs, among which amino acid residue 893 and 894 had a statistical difference with gastric cancer (P=0.004).

Conclusions: In this study, 503 CagA sequences was studied and analyzed in depth. In Chinese population, most *H. pylori* isolates are of the CagA-ABD subtype and its presence was associated with gastroduodenal disease. Amino acid polymorphisms at residue 893 and 894 flanking the EPIYA motif had a statistically significant association with gastric cancer.

Background

Helicobacter pylori (*H. pylori*) is a spiral, microaerophilic gram-negative bacterium that colonizes the gastric mucosa of more than half of the worldwide population [1]. *H. pylori* infection is not only closely related to chronic gastritis (CG) and peptic ulcer disease (PUD) but also an important risk factor for gastric cancer (GC) and mucosal-associated lymphoid tissue (MALT) lymphoma. Therefore, the World Health Organization classified *H. pylori* as a group carcinogen in 1994 [2–4]. Epidemiological survey shows that about 50% of adults in developed countries are infected with *H. pylori*, while the infection rate in developing countries is as high as 90% [5]. Despite the high prevalence of *H. pylori* infection, more than 80% of the carriers present asymptomatic gastritis, only 10%-20% develop CG and PUD, and a minority of *H. pylori* carriers develop into GC [6, 7]. Variation in virulence of the strains is thought to be an important reason for the different clinical outcomes of *H. pylori* infection [8]. Cytotoxin-associated gene A (*cagA*) is one of the most important virulence genes of *H. pylori*, which is located at the end of cag pathogenicity island (cag PAI) and encodes the 120–145 kDa CagA protein. Studies have confirmed that the *cagA*-positive strains are more virulent than the *cagA*-negative strains and can cause more severe gastric inflammation. CagA protein can be transported into the gastric epithelial cells by type IV secretion system (T4SS) encoded by the cag PAI. After the CagA translocation, the tyrosine residues of EPIYA(Glu-Pro-Ile-Tyr-Ala) motif at the 3' variable region can be phosphorylated by Src family kinases (SFKs) rapidly [9]. Based on the amino acid sequences polymorphisms, CagA C-terminal variable region can be divided into four different segments: EPIYA-A, EPIYA-B, EPIYA-C and EPIYA-D [10]. According to the different combinations of these four EPIYA motifs, *H. pylori* can be divided into two types, namely the East Asian type and the Western type [11].

CagA can specifically bind to the SH2 domain of Src homology 2 (SH2) - containing protein tyrosine phosphatase (SHP-2), which induces spatial configuration change of SHP-2 and activates it [12]. SHP-2 can be involved in the downstream signal transduction of growth factor receptor, regulate cell growth, differentiation and cell adhesion, and thereby inducing morphologic transformation and abnormal proliferation of gastric epithelial cells [13]. The binding of CagA and SHP-2 can lead to the cytoskeletal rearrangement of the host gastric epithelial cells, known as the hummingbird phenotype, which plays an important role in the development of gastric cancer [14]. Studies showed that the East Asian type CagA containing EPIYA-D segment displayed stronger binding activity to SHP-2 and more strongly damage to cells than does Western CagA. Western strains with more EPIYA-C segments show a stronger ability to bind to SHP2 and can be prone to induce the hummingbird phenotype than Western type CagA containing segments EPIYA-C [15, 16].

The incidence of *H. pylori* infection and gastric cancer in China is much higher than that in the western countries [17]. However, there are controversial reports about the relationship between the CagA type and gastroduodenal disease. This controversy may be due to regional diversity or differences in research methods. In fact, there is lack of comprehensive analysis of CagA 3' variable region sequence characteristics. Moreover, few studies have detected the detailed amino acid polymorphisms surrounding the EPIYA motif and their

association with clinical outcomes [18]. The aim of this study was to investigate the diversity of CagA 3' variable region and the amino acid polymorphisms surrounding the EPIYA motif, and the relationship with gastroduodenal disease through the sequence alignment and statistical analysis of 503 CagAs in *H. pylori* strains isolated from Chinese population.

Results

cagA gene status

A total of 503 (97.7%) *cagA*-positive strains out of 515 *H. pylori* isolates from 14 different geographical regions in China were obtained. Among those *cagA*-positive strains, 82(91.1%) were isolated from Shandong, 75 (94.9%) from Guangxi, 100% from other twelve regions. There was no difference in the distribution among different regions ($\chi 2 = 0.933$, $P > 0.05$). All 131 strains with clinical information were positive for the *cagA* gene.

Characteristics of EPIYA segments flanking sequences

The sequence of CagA 3'variable region begins with five polyaspartic amide (NNNNN) and ends with LSKVG. According to the segments flanking EPIYA motifs, we classified EPIYA segments. In addition to the four major segments, we defined several rare segments, including EPIYA-B', EPIYA-B" and EPIYA-D'. Representative segment types obtained from 503 CagAs were listed in Table 1. Through sequence alignment, it was found that there were differences in amino acids among the same sequences. The two most frequent segments in segments A, B, C and D are shown in Table 2.

Table 1

Representative sequences of EPIYA repeat region*

Segment	No.	Representative sequences
A_D	483	KELNEKLFGNSNNNNNGLKNNT**EPIYA**QVNKKK
B_D	478	TGQVASPE**EPIYA**QVAKKVSAKIDQLNEATS
B'_D	56	TGQVASPE**EPIYA**QVNKKK
B''_D	18	AINRKIDRINKIASAGKGVGGFSGAGQSASPE**EPIYA**QVAKKVSAKIDQLNESAS
D	468	AINRKIDRINKIASAGKGVGGFSGAGRSASP**EPIYA**TIDFDEAN
D'	12	FPLKRHDKVGDLSKVGLSASP**EPIYA**TIDFDEAN
A_C	22	KELNEKFKNFNNNNNGLKN**EPIYA**KVNKKK
B_C	22	TGQVASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG
C	28	FPLKRHDKVDDLSKVGLSASP**EPIYA**TIDDLGGP

*The subscripts C and D indicate that sequences containing segments A, B, B' and B" contain segments C and D, respectively.

Table 2
Two most frequent segments in EPIYA repeat region *

Segment	Sequences	Ratio
A_D	KELNEKLFGNSNNNNNGLKNNT**EPIYAQ**VNKKK	145/483
A_D	KELNEKLFGNSNNNNNGLKNNT**EPIYAK**VNKKK	29/483
B_D	TGQ**VA**SPE**EPIYA**QVAKKVSAKIDQLNEATS	98/478
B_D	TGQ**AT**SPE**EPIYA**QVAKKVSAKIDQLNEATS	89/478
D	AINRKIDRINKIASAGKGVGGFSGAG**R**SASP**EPIYA**TIDFDEAN	179/468
D	AINRKIDRINKIASAGKGVGGFSGAG**Q**SASP**EPIYA**TIDFDEAN	85/468
A_C	KELN**E**K**FK**NFNNNNNGLKN**EPIYA**KVNKKK	15/22
A_C	KELN**A**K**LG**NFNNNNNGLKN**EPIYA**KVNKKK	6/22
B_C	TGQ**V**ASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG	5/22
B_C	TGQ**A**ASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG	2/22
C	FPLKRHDKVDDLSKVG**L**SASP**EPIYA**TIDDLGGP	16/28
C	FPLKRHDKVDDLSKVG**R**SVSP**EPIYA**TIDDLGGP	4/28

*Different amino acids in the two sequences are highlighted; Ratio = (Number of the segment)/(Total).

The alignment of the amino acid sequences confirmed that the EPIYA motifs in the EPIYA-A, EPIYA-C and EPIYA-D segments were relatively conservative, whereas EPIYA motif variation had the highest frequency in EPIYA-B segment. A total of 1,587 EPIYA motifs were obtained from the 503 CagAs, including 12 types of EPIYA or EPIYA-like sequences (Table 3). For the three most frequent motifs (excluding EPIYA), 73 out of 74 EPIYTs, 22 out of 23 ESIYAs and all 9 EPLYAs, appeared in segment B. As is shown in Table 4, the sequences, KVNK and QVNK, were the main types of A_C and A_D, respectively. QVAK was the main amino acid of segments B_C and B_D. In the present study, the sequences are identified as EPIYA segments C and D if they are followed by TIDD and TIDF, respectively. However, by sequence alignment, it also belongs to segment C if it is followed by TIED or TIDE.

Table 3
Distribution of EPIYA motifs in segments A, B, C and D

Type	Distribution	Total
A	502 EPIYA, 1 ESIYA, 1 EPVYA, 1 EPIYT	505
B, B', B"	450 EPIYA, 73 EPIYT, 22 ESIYA, 9 EPLYA, 7 ESIYT, 5 ELIYA, 3 EHIYA, 1 EAIYA, 1 APIYA, 1 ELIYA, 1 DPIYA	573
D, D'	480 EPIYA	480
C	29 EPIYA	29
Total		1587

Table 4
Distribution of the first four amino acids following EPIYA motifs

Type	EPIYA-D		EPIYA-C	Total
	Occurrence and short segment	Total	Occurrence and short segment	
A	356 QVNK, 119KVNK, 4 EVNK, 4 QVAK	483	16 KVNK, 6 EVNK	22
B	477 QVAK, 60 QVNK, 4 KVNK, 3 QIAK, 2 QVTK, 2 QVAR, 1 QLTK, 1 QITK, 1 QVAQ, 1QVNG	552	22 QVAK	22
C/D	480 TIDF	480	24 TIDD, 2 TIDE, 2 TIED	28

CagA sequence type classification

A total of 20 sequence types were obtained from 503 CagAs (Table 5). CagA type was mainly East Asian type, accounting for 95.6% (481/503). The majority of the sequences were of types ABD (82.1%, 413/503) and AB'BD (8.2%, 41/503). There were only 22 strains of Western type, including types AC, ABC, ABCC and ABCCCC. There were 1–8 EPIYA motifs in CagA C-terminal variable region, and 87.3% (439/503) of the strain sequences had three EPIYA segments. The sequences containing 1 through 9 EPIYA segments are 1, 6, 439, 46, 6, 2, 2, 1 and 0, respectively. For example, there was only one EPIYA segment D in the sequence of type D and eight EPIYA segment in the sequence of AB'B'B'B'B'BD, including six repeats of segment B.

Table 5

Number of the sequence types*

SEq. Type	No.	SEq. Type	No.	SEq. Type	No.	SEq. Type	No.
ABD	413	AAABD	1	BD	1	D	1
ABD'	5	ABDABD	1	AB'BB"DAB'	1	ABC	16
AB-D'	5	AD	2	AB'B'B'B'BD	1	ABCC	4
AB'BD	41	AD'	1	ABB"B"	1	ABCCCC	1
AB'B'BD	5	A-D'	1	AB'B'B'B'B'BD	1	AC	1

*The hyphen indicates that there is no EPIYA motif between two adjacent EPIYA segments.

Correlation between CagA sequence types and geographical regions

There are some differences in CagA sequence types in different geographic regions (Table 6). In Yunnan, strains containing 4 or more EPIYA motifs accounted for 40% (29/73). There was a significant correlation between CagA-AB'BD type and Yunnan isolates ($\chi^2 = 81.523$, $P < 0.001$). However, most of the Western strains were from Neimenggu, and there was a significant difference between CagA-ABC and Neimenggu isolates ($\chi^2 = 25.468$, $P < 0.01$). The CagA-ABD type had statistical difference in different geographic regions ($\chi^2 = 80.067$, $P < 0.01$).

Table 6

The distribution of CagA sequence types in different geographic regions*

EPIYA type	CagA type	geographic areas														Total
		FJ	SD	GX	YN	HL	HN	NM	QH	ZJ	BJ	NX	TW	SX	XZ	
East Asian type	ABD	65	75	69	41	49	46	18	32	6	4	3	4	1	0	413
	ABD'	0	1	3	0	0	0	1	0	0	0	0	0	0	0	5
	AB-D'	0	0	0	0	0	0	5	0	0	0	0	0	0	0	5
	AB'BD	10	3	1	24	1	0	0	0	0	0	1	0	0	1	41
	AB'B'BD	2	0	0	3	0	0	0	0	0	0	0	0	0	0	5
	AAABD	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	ABDABD	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	AD	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
	AD'	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	A-D'	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	BD	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
	AB'BB"DAB'	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	AB'B'B'B'BD	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	ABB"B"	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	AB'B'B'B'BD	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	D	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
western type	ABC	0	2	1	2	5	0	5	1	0	0	0	0	0	0	16
	ABCC	0	0	1	1	0	0	2	0	0	0	0	0	0	0	4
	ABCCCC	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
	AC	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Total	83	82	75	73	56	46	33	33	7	4	4	4	2	1	503

*Isolates were from 14 regions: Fujian (FJ), Shandong (SD), Guangxi (GX), Yunnan (YN), Heilongjiang (HL), Hunan (HN), Neimenggu (NM), Qinghai (QH), Zhejiang (ZJ), Beijing (BJ), Ningxia (NX), Taiwan (TW), Shanxi (SX) and Xizang (XZ).

Correlation between CagA sequence types and clinical outcomes

In all 131 samples, 64.9%, 16.8%, 7.6%, 7.6% and 3.1% of the patients had diseases CG, GC, GU, DU and MALT, respectively, which shows that gastritis patients accounted for the majority of the proportion in these samples. A total of 12 Western type strains were found, 11 of which were from patients with CG. Among all the 131 CagAs, 86 were of the type ABD, 25 of the type AB'BD, 3 of the type AB'B'BD, 2 of the type AB'B'B'BD and 3 of the type AD. The distribution of the CagA sequence types in various clinical outcomes is shown in Table 7. There were significantly statistical difference between the ABD/AB'BD subtypes and clinical diseases (χ^2 = 80.067/71.500, P < 0.01). As is shown in Table 7, the prevalence of ABD was 58.1% (50/86) in G; whereas only 22.1% (19/86) in GC and 9.3% (8/86) in GU. The ratio of AB'BD /ABD was therefore higher in CG (20/50 = 0.4) than GC (1/19 = 0.05).

Table 7
CagA sequence types and clinical outcomes

	ABD	AB'BD	AB'B'BD	AB'B'B'B'BD	AD	AC	ABC	ABCC	ABCCCC	Total (%)
CG	50	20	2	1	1	1	6	3	1	85 (64.9)
GC	19	1	0	1	1	0	0	0	0	22 (16.8)
GU	8	1	0	0	1	0	0	0	0	10 (7.6)
DU	5	3	1	0	0	0	1	0	0	10 (7.6)
MALT	4	0	0	0	0	0	0	0	0	4 (3.1)
Total	86	25	3	2	3	1	7	3	1	131 (100)

Amino acid polymorphisms flanking the EPIYA motifs

Sequence alignment analysis showed that there were amino acid polymorphisms flanking the EPIYA motifs of *H. pylori* CagA 3' variable region. There were seven amino acid polymorphisms, at residues 893, 894, 900, 906, 909, 910 and 963, where the substitution rate of amino acids was more than 18.6% ([5 + 11]/ [70 + 5 + 11]). The detailed information of these amino acid polymorphisms in the sequence flanking the EPIYA motif in 86 ABD subtypes are shown in Table 8. The absence of amino acids 893 and 894 was synchronous. Strains at the absence of the 893 and 894 residues had a statistically significant association with GC compared with CG ($\chi^2 = 21.778$, $P < 0.01$). Most patients with CG, GU, DU and MALT had a glutamic acid (Glu) at 894, while some patients with GC had Glu deletion or substituted by other amino acids, such as threonine (Thr) or asparagine (Asn). These changes at residue 894 had significant difference between GC patients and those with other diseases ($\chi^2 = 4.908$, $P < 0.05$). In addition to the seven amino acid polymorphisms mentioned above, other amino acids were relatively conservative, except for individual amino acid absence or substitution.

Table 8
Amino acid polymorphisms in the sequence flanking the EPIYA motif in 86 ABD subtypes*

Disease	893			894				900			906		909		910			963			Total
	N	S	~	E	T	~	A/N	Q	K	E	T	A	A	V	A	T	V/I	R	Q	L	
G	43	3	4	32	12	4	2	29	20	1	32	18	28	22	34	15	1	34	15	1	50
GC	12	0	7	2	9	7	1	14	5	0	10	9	10	9	12	7	0	12	7	0	19
GU	8	0	0	5	2	0	1	8	0	0	5	3	6	2	5	3	0	7	1	0	8
DU	4	1	0	4	0	0	1	5	0	0	2	3	3	2	4	1	0	3	2	0	5
MALT	3	1	0	3	1	0	0	2	2	0	2	2	3	1	2	1	1	3	1	0	4
Total	70	5	11	46	24	11	5	58	27	1	51	35	50	36	57	27	2	59	26	1	86

*~: amino acid deletion at the position; N: asparagine; S: serine; E: glutamic acid; T: threonine; A: alanine; Q: glutamine; K: lysine; V: valine; I: isoleucine; R: arginine; L: Leucine

Discussion

CagA is an important oncoprotein that can be translated into the gastric epithelial cells and subsequently tyrosine-phosphorylated at residues of the EPIYA motifs. The phosphorylated CagA can activate the phosphatase SHP-2 and then cause actin cytoskeleton rearrangement, hummingbird phenotype, which disturbs the normal signal transduction pathway of cells and promotes abnormal proliferation of gastric epithelial cells. The interaction between CagA and SHP-2 suggests that CagA, as a key oncoprotein, plays an important role in the development of gastrointestinal diseases caused by *H. pylori*. Therefore, we used molecular epidemiological methods to study the diversity of CagA 3' variable region and explore the molecular mechanisms by which *H. pylori* infection promotes the development of gastrointestinal diseases.

The tyrosine phosphorylation site is located on EPIYA repeat sequences at the CagA C-terminus, and the number of EPIYA repeats directly affects the binding of CagA to SHP-2 and the ability of causing morphological changes of gastric epithelial cells [20]. Therefore, the variation of EPIYA repeat sequences may be an important reason for the difference in *H. pylori* strains virulence and clinical outcome. In

our study, the EPIYA motifs variation in EPIYA-B segments were more frequent than in the EPIYA-A, EPIYA-C and EPIYA-D segments, and the EPIYT (74/1,587) was the most common variant type. In EPIYA-C and EPIYA-D segments, the amino acids following EPIYA motif are generally TIDD and TIDF, respectively, which is an important structural domain of binding SHP-2. Our study confirms that the EPIYA belongs to segment C if it is followed by TIED or TIDE. However, it has been proven that EPIYA is also identified as segment C if it is followed by TIEE, SIDD, TIDG, TIAE or TIAD, and it belongs to segment D if followed by TIDS [2].

According to the segments flanking the EPIYA motifs, we defined several segments, including B'_D, B''_D and D'. The sequences of B'_D, B''_D and D' segments have some differences from those of B and D segments. For example, the sequences before EPIYA are similar to those of D segment in B''_D segment, whereas the sequences after EPIYA are similar to those of A_D segment in B'_D segment. It has been reported that the distribution of CagA EPIYA segments shows great geographical differences. The EPIYA-A and EPIYA-B segments appear in almost all *cagA*-positive strains, whereas EPIYA-C and EPIYA-D segments are characteristic of Western and East Asian CagA strains, respectively. In our study, 82.1% (413/503) of the CagA strains were of the ABD subtype, whereas 4.0% (20/503) were of the ABC or ABCC subtype. 77.3% (17/22) of the Western CagAs were from Neimenggu, Heilongjiang and Yunnan, which may be due to human migration or direct transmission. Studies have reported that there was no significant correlation between CagA-ABD and the types of gastroduodenal diseases [21, 22]. However, our study confirmed that there was significant correlation between the ABD subtype and gastroduodenal diseases ($P < 0.01$). Studies have shown that East Asian CagA is more pathogenic than Western Cag A, which may explain why the incidence of gastric cancer in eastern countries is significantly higher than that in western countries [23, 24].

CagA can be phosphorylated by the SFKs at tyrosine residues of the EPIYA motifs [25]. The tyrosine phosphorylated C and D segments specifically bind to SHP2, which plays an important role in the development of gastric cancer [26, 27]. The tyrosine phosphorylated A and B segments can bind and activate the CagA C-terminal Src kinase (CSK) that is a SFK with negative feedback regulation [28, 29]. The inhibition of SFK can lead to the decrease of phosphorylated CagA protein, which to some extent explains that *H. pylori* can survive in gastric epithelial cells for a long time without causing extensive gastric injury [29]. Therefore, it is thought that CagA with more A and B segments can inhibit SFK more effectively, and thereby reduce cell damage [30, 31]. In the present study, we found 20 CagA sequence types with different numbers of the EPIYA-A or EPIYA-B segment, such as AAABD, ABDABD and BD. The number of EPIYA-A and EPIYA-B segments may lead to the difference in the type and severity of gastrointestinal diseases. The relationship between EPIYA segments and gastrointestinal diseases needs to be further explored.

Research has shown that the pathogenicity of CagA is determined by the binding ability of SHP-2, which is also related to the number of tyrosine phosphorylation sites [12]. Souza [32] reported that the SH2 domains bound to highly correlated sequences, and the binding motif is pY-(S/T/A/V/I)-X-(V/I/L)-X-(W/F). Interestingly, the binding ability of East Asian CagA (pY-A-T-I-D-F) to SHP-2 is higher than that of Western CagA (pY-A-T-I-D-D), which can lead to more severe gastroduodenal diseases. Higashi et al. [10] demonstrated that the difference of single amino acid led to the difference of SHP-2 binding activity between East Asian and Western CagA proteins. Therefore, the research on amino acid polymorphisms and their assosiation with gastrointestinal diseases may have an important clinical value. In our study, we obtained seven amino acid polymorphisms in the sequences surrounding the EPIYA motifs: residues 893, 894, 900, 906, 909, 910 and 963. The absence of the amino acid 893 and 894 had a statistically significant association with GC. In most patients with CG, GU, DU and MALT, the amino acids at residues 893 and 894 were asparagine (Asn) and glutamic acid (Glu), respectively, whereas 36.8% (7/19) of the isolates from GC patients lost these two amino acids. This change may affect the ability of CagA tyrosine phosphorylation and binding to SHP-2, and alter the spatial conformation of CagA protein, thereby accelerating the development of gastrointestinal diseases.

Conclusions

In this study, 503 CagA sequences were analyzed in depth and we defined several novel segment types, including B'_D, B''_D and D'. We demonstrated that most of *H. pylori* isolates from Chinese population were of the CagA-ABD subtype and it was statistically correlated with the type of gastroduodenal diseases. Strains at the absence or mutation of the 893 and 894 residues had a significant association with GC. Therefore, amino acid polymorphism in EPIYA motifs might affect the function of CagA protein, and then lead to the development of gastrointestinal diseases, especially GC.

Methods

H. pylori culture and DNA extraction

A total of 515 *H. pylori* isolates preserved in our laboratory were obtained from the following regions: Fujian (n = 83), Shangdong (n = 90), Guangxi (n = 79), Yunnan (n = 73), Heilongjiang (n = 56), Hunan (n = 46), Neimenggu (n = 33), Qinghai (n = 33), Zhejiang (n = 7), Beijing (n = 4), Ningxia (n = 4), Taiwan (n = 4), Shanxi (n = 2), Xizang (n = 1). *H. pylori* were grown on 5% defibrinated sheep blood agar plates at 37 °C for 3–5 days in a microaerobic atmosphere (5% O_2, 10% CO_2 and 85% N_2). Bacteria were identified as *H. pylori* based on its external morphology, negative Gram staining and positive for catalase, oxidase and urease. The confirmed isolates were frozen at -80 °C until the genomic DNA was extracted with the QIAamp DNA Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The extracted DNA was stored at -20°C and used directly for PCR. Due to the incomplete clinical data collected, we obtained clinical information for 131 strains out of the 509 *H. pylori* isolates in the present study. Based on the gastrointestinal endoscopy and pathological examination, CG was diagnosed in 85 patients, GC in 22, gastric ulcer (GU) in 10, duodenal ulcer (DU) in 10 and MALT lymphoma in 4. This study was approved by the Research and Ethical committees.

PCR amplification

To amplify the *cagA* 3' variable region of *H. pylori*, the primers were: forward, 5'- ATAATGCTAAATTAGACAACTTGAGCGA – 3' and reverse, 5'- TTAGAATAATCAACAAACATCACGCCAT – 3' with a 297-bp product [19]. All PCR were performed in a volume of 25 mL containing 25 μl containing 1 μl each of primer, 1 μl template DNA, 12.5 μl Go Taq® Green Master Mix (Promega, USA) and 9.5 μl nuclease-free water. PCR was performed using a thermocycler system (Bio-Rad, USA) under the following conditions: denaturation at 94 °C for 5 min, 35 cycles at 94 °C for 30 s, at 54 °C for 30 s and at 72 °C for 40 s, and an extension at 72 °C for 10 min. The amplified products were identified after electrophoresis on 1.5% agarose gels with GelStain in 1 × TAE buffer at 110 V for 30 min. The gel documentation system (Bio-Rad, USA) was used to detect the DNA bands and obtain the images of the PCR products.

Sequencing and analysis of the diversity of the CagA 3' variable region

Positive PCR products were sent to the Beijing Genomics Institute (BGI) for purification and sequencing. EditPlus (version 5.3.0, korea) was used to collect sequence information, sort the sequences and create files in FASTA format. Bioedit was used to align and obtain amino acid sequences of the CagA protein. The Western strain 26695 *cagA* (GenBank No. CP003904) was used as a reference sequence. MEGA software (version 7.0.18, USA) was used for sequence alignments to analyze the diversity of the CagA 3' variable region.

Statistics

Statistical datas were analyzed using SPSS 20.0 (SPSS, Chicago, USA). The χ^2 test and Fisher's exact test were used to test statistical difference among different gastroduodenal diseases in the CagA subtype and amino acid polymorphisms. A P-value < 0.05 was considered indicative of a statistically difference.

Abbreviations

H.pylori: *Helicobacter pylori*

EPIYA: Glu-Pro-Ile-Tyr-Ala

PCR: polymerase chain reaction

CG: chronic gastritis

PUD: peptic ulcer disease

GC: gastric cancer

MALT: mucosal-associated lymphoid tissue

GU: gastric ulcer

DU: duodenal ulcer

cagA: cytotoxin-associated gene A

cag PAI: cag pathogenicity island

T4SS: type IV secretion system

SFKs: Src family kinases

SHP-2: Src homology 2 (SH2) - containing protein tyrosine phosphatase

CSK: CagA C-terminal Src kinase

BGI: Beijing Genomics Institute

Declarations

Availability of data and materials

The Western strain 26695 *cagA* was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/).

Acknowledgements

Not applicable.

Funding

This work was supported by the National Science and Technology Major Project of China (2018ZX10712-001).

Ethics approval and consent to participate

Not applicable

Consent for publication

Full consent is given for publication in Gut Pathogens.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZX developed the idea, designed the study, collected the samples, analyzed the datas and drafted the manuscript. YY and LH performed the DNA extraction. YG, XH and RF collected the samples. MZ and XY designed the study and analyzed the results. JZ designed the study, reviewed and revised the manuscript. All authors read and approved the final manuscript.

References

1. Wen S, Moss SF. *Helicobacter pylori* virulence factors in gastric carcinogenesis. Cancer Lett. 2009;282(1):1–8.
2. Atherton JC. The pathogenesis of *Helicobacter pylori*-induced gastro-duodenal diseases. Annu Rev Pathol. 2006;1(1):63–96.
3. Suerbaum S, Michetti P. *Helicobacter pylori* Infection. New Engl J Med. 2002;347(15):1175–86.
4. Blaser MJ. *Helicobacter pylori* and gastric diseases. BMJ. 1998;316(7):1507–10.
5. Cover TL, Dooley CP, Blaser MJ. Characterization of and human serologic response to proteins in *Helicobacter pylori* broth culture supernatants with vacuolizing cytotoxin activity. Infect Immun. 1990;58(3):603–10.
6. Ogorodnik E, Raffaniello RD. Analysis of the 3'-variable region of the *cagA* gene from *Helicobacter pylori* strains infecting patients at New York City hospitals. Microb Pathog. 2013;56(2):29–34.

7. Chomvarin C, Phusri K, Sawadpanich K, Mairiang P, Hahnvajanawong C. Prevalence of CagA EPIYA motifs in *Helicobacter pylori* among dyspeptic patients in northeast Thailand. Southeast Asian J Trop Med Public Health. 2012;43(1):105–15.
8. Tummuru MKR, Cover TL, Blaser MJ. Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori*: Evidence of linkage to cytotoxin production. Infect Immun. 1993;61(5):1799–09.
9. Odenbreit S, Püls J, Sedlmaier B, Gerland E, Fischer W, Haas R. Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. Science. 2000;287(5457):1497–500.
10. Higashi H, Tsutsumi R, Fujita A, Yamazaki S, Asaka M, Azuma T, et al. Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. Proc Natl Acad Sci USA. 2002;99(22):14428–33.
11. Hatakeyama M. Oncogenic mechanisms of the *Helicobacter pylori* CagA protein. Nat Rev Cancer. 2004;4(9):688–94.
12. Higashi H. SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein. Sci. 2002;295(5555):683–6.
13. Selbach M, Moese S, Hauck CR, Meyer TF, Backert S. Src is the kinase of the *Helicobacter pylori* CagA protein in vitro and in vivo. J Biol Chem. 2002;277(9):6775–8.
14. Tammer I, Brandt S, Hartig R, König W, Backert S. Activation of Abl by *Helicobacter pylori*: a novel kinase for *cagA* and crucial mediator of host cell scattering. Gastroenterology. 2007;132(4):1309–19.
15. Ferreira RM, Machado JC, Leite M, Carneiro F, Figueiredo C. The number of *Helicobacter pylori* CagA EPIYA-C tyrosine phosphorylation motifs influences the pattern of gastritis and the development of gastric carcinoma. Histopathology. 2012;60(6):992–8.
16. Souza DD, Fabri LJ, Nash A, Hilton DJ, Baca M. SH2 domains from suppressor of cytokine signaling-3 and protein tyrosine phosphatase SHP-2 have similar binding specificities. Biochemistry. 2002;41(29):9229–36.
17. Figueroa G, Troncoso M, Toledo MS, Faúndez G, Acua R. Prevalence of serum antibodies to *Helicobacter pylori vacA* and *cagA* and gastric diseases in Chile. J Med Microbiol. 2002;51(4):300–4.
18. Argent RH, Kidd M, Owen RJ, Thomas RJ, Limb MC, Atherton JC. Determinants and consequences of different levels of CagA phosphorylation for clinical isolates of *Helicobacter pylori*. Gastroenterology. 2004;127(2):514–23.
19. Gerhard M, Lehn N, Neumayer N. Clinical relevance of the *Helicobacter pylori* gene for blood-group antigen-binding adhesin. Proc Natl Acad Sci USA. 1999;96(22):12778–83.
20. Higashi H, Yokoyama K, Fujii Y, Ren S, Yuasa H, Saadat I, et al. EPIYA motif is a membrane-targeting signal of *Helicobacter pylori* virulence factor CagA in mammalian cells. J Biol Chem. 2005;280(24):23130–8.
21. Zhou J. *CagA* genotype and variants in Chinese *Helicobacter pylori* strains and relationship to gastroduodenal diseases. J Med Microbiol. 2004;53(3):231–5.
22. Chen CY, Wang FY, Wan HJ, Jin XX, Wei J, Wang ZK, et al. Amino acid polymorphisms flanking the EPIYA-A motif of *Helicobacter pylori* CagA C-terminal region is associated with gastric cancer in east China: experience from a single center. J Dig Dis. 2013;14(7):358–65.
23. Miura M, Ohnishi N, Tanaka S, Yanagiya K, Hatakeyama M. Differential oncogenic potential of geographically distinct *Helicobacter pylori cagA* isoforms in mice. Int J Cancer. 2009;125(11):2497–504.
24. Satomi S, Yamakawa A, Matsunaga S, Masaki R, Inagaki T, Okuda T, et al. Relationship between the diversity of the *cagA* gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan. J Gastroenterol. 2006;41(7):668–73.
25. Odenbreit S. Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. Science. 2000;287(5457):1497–500.
26. Miyamoto D, Miyamoto M, Takahashi A, Yomogita Y, Higashi H, Kondo S, et al. Isolation of a distinct class of gain-of-function SHP-2 mutants with oncogenic RAS-like transforming activity from solid tumors. Oncogene. 2008;27(25):3508–15.
27. Hatakeyama M, Higashi H. *Helicobacter pylori cagA*: a new paradigm for bacterial carcinogenesis. Cancer Sci. 2005;96(5):835–43.
28. Higashi H. SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein. Science. 2002;295(5555):683–6.
29. Tsutsumi R, Higashi H, Higuchi M, Okada M, Hatakeyama M. Attenuation of *Helicobacter pylori* CagA·SHP-2 signaling by interaction between CagA and C-terminal Src kinase. J Biol Chem. 2003;278(35):3664–70.
30. Hatakeyama M. Anthropological and clinical implications for the structural diversity of the *Helicobacter pylori* CagA oncoprotein. Cancer Sci. 2011;102(1):36–43.
31. Furuta Y, Yahara K, Hatakeyama M, Kobayashi I. Evolution of *cagA* Oncogene of *Helicobacter pylori* through Recombination. Plos One. 2011;6(8):e23499.

32. Souza DD, Fabri LJ, Nash A, Hilton DJ, Baca M. SH2 domains from suppressor of cytokine signaling-3 and protein tyrosine phosphatase SHP-2 have similar binding specificities. Biochemistry. 2002;41(29):9229–36.