OPEN ACCESS Freely available online

PLoS one

A Comprehensive Sequence and Disease Correlation Analyses for the C-Terminal Region of CagA Protein of *Helicobacter pylori*

Youlin Xia[1], Yoshio Yamaoka[2,3]*, Qi Zhu[1], Ivan Matha[1], Xiaolian Gao[1]*

1 Department of Biology and Biochemistry, University of Houston, Houston, Texas, United States of America, **2** Department of Medicine-Gastroenterology, Veterans Affairs Medical Center and Baylor College of Medicine, Houston, Texas, United States of America, **3** Department of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Yufu, Japan

Abstract

Chronic *Helicobacter pylori* infection is known to be associated with the development of peptic ulcer, gastric cancer and gastric lymphoma. Currently, the bacterial factors of *H. pylori* are reported to be important in the development of gastroduodenal diseases. CagA protein, encoded by the *cagA*, is the best studied virulence factor of *H. pylori*. The pathogenic CagA protein contains a highly polymorphic Glu-Pro-Ile-Tyr-Ala (EPIYA) repeat region in the C-terminal. This repeat region is reported to be involved in the pathogenesis of gastroduodenal diseases. The segments containing EPIYA motifs have been designated as segments A, B, C, and D; however the classification and disease relation are still unclear. This study used 560 unique CagA sequences containing 1,796 EPIYA motifs collected from public resources, including 274 Western and 286 East Asian strains with clinical data obtained from 433 entries. Fifteen types of EPIYA or EPIYA-like sequences are defined. In addition to four previously reported major segment types, several minor segment types (e.g., segment B′, B′′) and more than 30 sequence types (e.g., ABC, ABD) were defined using our classification method. We confirm that the sequences from Western and East Asian strains contain segment C and D, respectively. We also confirm that strains with two EPIYA segment C have a greater chance of developing gastric cancer than those with one segment C. Our results shed light on the relationships between the types of CagAs, the country of origin of each sequence type, and the frequency of gastric disease.

Citation: Xia Y, Yamaoka Y, Zhu Q, Matha I, Gao X (2009) A Comprehensive Sequence and Disease Correlation Analyses for the C-Terminal Region of CagA Protein of *Helicobacter pylori*. PLoS ONE 4(11): e7736. doi:10.1371/journal.pone.0007736

Editor: Niyaz Ahmed, University of Hyderabad, India

Received July 18, 2009; **Accepted** September 15, 2009; **Published** November 6, 2009

Funding: This research was in part supported by grants from NIH (R42 GM067364 to XG and RO1 DK62813 to YY) and the Robert A. Welch Foundation (E-1027 to XG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yyamaoka@bcm.tmc.edu (YY); xgao@uh.edu (XG)

Introduction

Helicobacter pylori is a Gram-negative bacterium etiologically involved in peptic ulcer disease, gastric adenocarcinoma, and primary gastric B-cell lymphoma [1]. Although infection with *H. pylori* almost always results in chronic active gastritis, only a fraction of those infected develop clinical disease. While this phenomenon remains unexplained, host genetics, host immune response, and the relationship of the host response to bacterial virulence factors are likely to be important factors. A tremendous number of groups have investigated the roles of putative virulence factors of *H. pylori*, and the best studied is the CagA protein [2–7]. CagA producing strains are reported to be associated with severe clinical outcomes, especially in Western countries [8–11].

CagA is a highly immunogenic protein with a molecular weight between 120 and 140 kDa [12,13]. Variation in the size of CagA is due to the presence of a variable number of repeat sequences located in the 3′ region of the gene [12,14–16]. The repeat regions contain the Glu-Pro-Ile-Tyr-Ala (EPIYA) motif. To characterize the different sequence patterns in the 3′ region, at least four methods of classification are typically reported. First, the terms D1, D2, and D3 are used to designate three specific sequences [12]. Second, sequences are denoted with combinations of R1, R2, and R3 [14,15]. Third, each EPIYA motif is assigned a motif type (e.g., EPIYA-A, -B, -C, or –D motif) [17,18,19]. Finally, sequences are annotated according to segments (20–50 amino acids) flanking the EPIYA motifs (segments EPIYA-A, -B, -C, or –D) [20–23], after the identification of the essential CagA phosphorylation sites as confirmed by mutagenesis during infection and transfection [24]. Initially, the two Csk binding sites are designated as segments EPIYA-A and –B, and the Src homology 2 (SH2) domain of Src homology 2 phosphatase (SHP-2) binding sites in Western and East Asian type CagA are designated as segments EPIYA-C and –D, respectively. Here, "motif" and "segment" are used to designate the five-member sequence (EPIYA) and the short sequences around the EPIYA motif, respectively (Figure 1). However, none of the four sequence classification methods work well with non-standard sequences, and a modified classification method was deemed necessary.

CagA is encoded by the *cagA* gene, which is located at one end of the *cag* pathogenicity island (PAI) [25]. The *cag* PAI encodes a type IV secretion system, by which CagA proteins are delivered into host cells [26–30]. CagA interacts with various target molecules in addition to Csk and SHP-2, including Src [31,32] and Abl [33]. Recent study clearly confirmed that almost one dozen of factors such as SHP-1, Grd2, Grb2, phosphatidylinositol 3-OH kinase

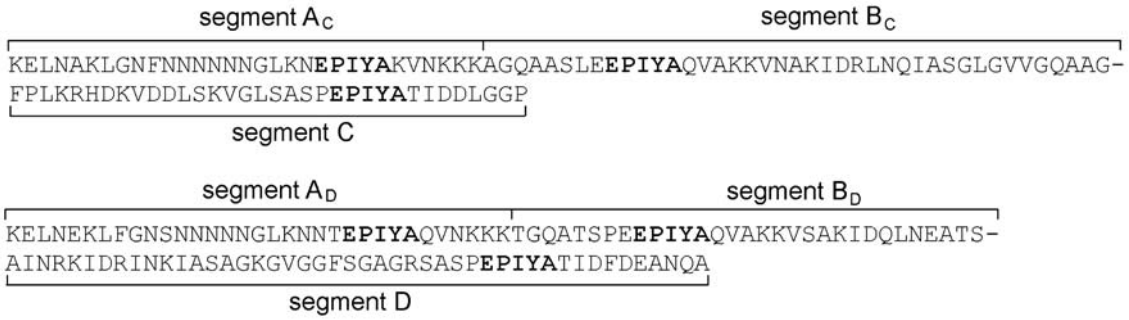

Figure 1. Definitions of segments around EPIYA motif (EPIYA or EPIYA-like sequences). The upper sequences are typical CagA sequences with Western type and the lower sequences are typical CagA sequences with East Asian type. Segments A, B, B′, and B′′have subscripts C and D, indicating that the sequences containing segments A, B, B′, and B′′ contain segments C and D, respectively. For example, the notation EPIYA-A_C signifies segment A from a CagA sequence containing the segment C.
doi:10.1371/journal.pone.0007736.g001

(PI3K), have also binding abilities to CagA phosphorylation sites [34]. Mutations of SHP-2 have been found in various human malignancies and altered SHP-2 signaling culminates in the development of gastric adenocarcinoma in genetically engineered mice [35,36], indicating that SHP-2 is involved in the development of gastric cancer. Recent studies reported that the East Asian type CagA containing segments EPIYA-D exhibits stronger binding activity for SHP-2 and a greater ability to induce morphological changes in epithelial cells than Western type CagA containing segments EPIYA-C [17,20,23]. The recent study also showed that *H. pylori* strains possessing East Asian type CagA have an ability to induce higher amounts of interleukin-8 from gastric epithelial cells than *H. pylori* strains possessing Western type CagA [37]. Accordingly, East Asian strains are believed to be more virulent than Western strains, and this might be the reason why the incidences of gastric cancer in East Asian countries are relatively higher than those in Europe, North America, and Australia (Data available at http://www-dep.iarc.fr/). In addition, the incidence of gastric cancer is reported to be higher in patients infected with strains carrying multiple EPIYA repeats compared to those infected with strains of a single repeat [14,15,38,39].

However, there are also controversial reports that the genotypes (DNA analysis) of the CagA repeat region are not associated with clinical outcomes [40–43]. This controversy might be due in part to the fact that genotypes are not necessarily mutations in protein sequences and that the previous studies of the diversity of CagAs and the relationship of diseases and protein sequence types used only limited information, mostly relying on their own data sets. Indeed, there lacked comprehensive study considering all CagAs deposited in GenBank (http://www.ncbi.nlm.nih.gov/). Moreover, although CagA EPIYA repeats can be assigned to consensus sequence types, the existing sequence analyses did not completely consider the sequence variation patterns in the CagA repeat region. An in-depth analysis of the non-typical type repeats [15,44] becomes necessary for addressing the question. In this study, we used sequence comparison and statistical method to analyze 560 unique CagAs selected from 4,534 CagAs from three data sources. Our results shed light on the relationships between the types of CagAs, the country of origin of each sequence type, and the frequency of gastric disease.

Results and Discussion

EPIYA Motifs Classification

By sequence alignment or pattern comparison, we found that there were sequences similar to EPIYA (such as EPIYT, ESIYT), although most sequences contained EPIYA. In this study, the EPIYA or EPIYA-like sequences were defined as any five member amino acid sequence with at least three amino acids corresponding to the sequence, EPIYA (where Y is always constant). By searching all sequences before data filtering, we obtained 16 types of EPIYA or EPIYA-like sequences. Of these, 15 types were chosen for further study because their surrounding sequences were similar to those of EPIYA (Table 1), indicating that these sequences might have a function similar to EPIYA. One sequence, MAIYA, from entry ABA26023 was excluded because the pattern of its flanking sequences was very different from those of the other 15 types of EPIYA or EPIYA-like sequences (Table 1). The 15 types listed in Table 1 are called EPIYA "motifs" for simplicity, in this work.

The frequency of each EPIYA motif in the filtered data set is listed in Table 1. In total, 1,796 EPIYA motifs were obtained from the 560 CagAs. On average, each CagA sequence contained approximately three EPIYA motifs. The three most frequent EPIYA motifs were EPIYA (1,657/1,796 = 92.3%), EPIYT (92/1,796 = 5.1%), and ESIYA (24/1,796 = 1.3%).

EPIYA Segments Classification

We categorized the EPIYA segments according to the segments flanking the EPIYA motifs (Figure 1). In addition to the four major segments originally designated, EPIYA-A, -B, -C, and –D [20,22], we designated several minor segments, including EPIYA-B′ and -B′′. Representative examples of these types of segments, derived from the 560 CagAs, are listed in Table 2 (a few more other types of segments with frequency less than 10 are given in Table S1. For simplicity, we refer to segment EPIYA-A, -B, -C, or –D as segment A, B, C, or D. Segments A, B, B′, and B′′ have subscripts C and D, which indicate that the sequences that contain segments A, B, B′, and B′′ contain segments C and D, respectively (Figure 1). However, 19 short sequences did not contain either segments C or D, and we manually assigned a subscript C or D to the segment type, according to their sequence patterns.

Table 1. Frequencies of the 15 types of EPIYA motifs.

Motif	EPIYA	EPIYT	ESIYA	ESIYT	EPIYV	EHIYA	ELIYA	EPVYA
Freq.	1657	92	24	7	3	2	2	2
Motif	EPIYD	EPIYS	EPKYA	EPRYA	ETIYA	KPIYA	NPIYA	Total
Freq.	1	1	1	1	1	1	1	1,796

doi:10.1371/journal.pone.0007736.t001

Table 2. Representative segments of EPIYA motifs[a].

Type	Freq.	Representative sequence
A_C	272	KELNAKLGNFNNNNNNGLKN..**EPIYA**KVNKKK
A_D	295	KELNEKLFGNSNNNNNGLKNNT**EPIYA**QVNKKK
B_C	262	TGQVASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG
B_D	281	TGQATSPE**EPIYA**QVAKKVSAKIDQLNEATS
C	343	FPLKRHDKVDDLSKVGRSVSP**EPIYA**TIDDLGGP
D	284	AINRKIDRINKIASAGKGVGGFSGAGRSASP**EPIYA**TIDFDEAN
B'_C	10	AGQAASPE**EPIYA**KVNKKK
B'_D	14	AGQATSPE**EPIYA**QVNKKK
B''_D	19	AINRKIDRINKIASAGKGVGGFSGAGRSANP**EPIYA**QVARKVSA-KIDQLNEATS
Total	1,780	

[a]Note: the values in the table are the frequencies of similar sequences, not the number of identical sequences within a sequence type. Other segments of 16 EPIYA motifs are listed in Table S1.

doi:10.1371/journal.pone.0007736.t002

We named the minor segments according to the patterns of the sections immediately following EPIYA (Table 2). This was because the four amino acids, TIDD and TIDF, following EPIYA in segments C and D, respectively are reported to be important for the binding of SHP-2 [17,24]. For example, segments B'_C and B'_D are shorter versions of segments B_C and B_D, respectively (Table 2). In segment B''_D, the sequences before EPIYA are similar to those of segment D, whereas the sequences after EPIYA are similar to those of segment B_D.

The segment B displayed the biggest change in the five amino acids; EPIYA motif (Table S2). For the three most frequent motifs (excluding EPIYA), 89 out of 92 EPIYTs, all 23 ESIYAs, and all 7 ESIYTs, appear in segment B. Interestingly, 88 EPIYT motifs belong to the segment B_C, and only 1 EPIYT belongs to the segment B_D. In contrast, the changes of the five amino acids in segments A, C, and D were relatively small. In other reports [18,19], the NPIYA, EPIYT, ESIYA and ESIYT motifs were named as A′, B′, B″ and B‴, respectively. However, their terminology seems to be confusing, otherwise all 15 types of pseudo EPIYA motifs should have different names. Their motif A′ belonged to our segment A and their B′, B″, and B‴ fell into our segments B, B′, or B′′ (Table S2).

CagA Sequence Type Classification

Each CagA sequence was assigned a sequence type consisting of the names of the EPIYA segments in its sequence (such as ABC or ABD) (Table S3). Depending on the number of EPIYA segments, they are termed as AnBnCn or AnBnDn, where "n" is the repeating motifs and does not have to be equal for A, B, C, and D types (e.g., ABCCCC). In the event that there was an additional segment that lacked an EPIYA motif between two neighboring EPIYA segments, a hyphen was added between the two EPIYA segments (e.g., A-C, A-D). In total there were 28 segments without EPIYA motifs between two neighboring EPIYA segments among the 560 CagAs (Table S3). These 28 interval segments are of various lengths and contents. In total, 41 different sequence types were found (Table S4). Among the 41 sequence types, 32 sequence types are remained (Table 3) after removing the types containing rare EPIYA segment types (i.e., B''_C, C′, D′, C′′ and D′′). The majority of the sequences were of types ABD (43%) and ABC (30%). Interestingly, there were no CagA sequences containing both segments C and D. This suggests hybridization (recombination) between Western and East Asian CagA is very rare.

A small number of CagAs were classified differently between our current study and previous studies (examples shown in Table S5. For example, the CagA sequence of BAF45291 was classified as AC in a previous study [44]. However the sequence type was A-C in our classification, which meant that an interval segment (VKAKIDQLNQAASGFGNVGQAG) lacking EPIYA-like motif was present between the sequences A and C. For the CagA sequence of BAF45283, the sequence type was reported to be

Table 3. Frequencies of the 32 sequence types[a].

Seq. Type	Freq.	Seq. Type	Freq.	Seq. Type	Freq.	Seq. Type	Freq.
ABD	240	AB′-ABD	4	C	2	ABCCCC	1
ABC	167	A-D	4	A	1	A-B″D	1
ABCC	51	A-ABD	3	AB′B′BC	1	AB-D	1
ABB″D	16	AB-ABD	2	ABB″BD	1	ABD-ABD	1
AB	15	AB′B′BD	2	AB′BCC	1	ABD-BD	1
ABCCC	10	AB′BD	2	AB′-C	1	ABD-D	1
AB′BC	6	ABCCCCC	2	AB-C	1	A-CCC	1
A-C	5	AB′D	2	ABCB″CC	1	CC	1

[a]All sequence types are listed in Table S3. Other sequence types are listed in Table S4.

doi:10.1371/journal.pone.0007736.t003

Table 4. Two most frequent EPIYA segments[a].

	Segment	Ratio
A_C	KELNAKLGNFNNNNNNGLKN..**EPIYA**KVNKKK	53/272
A_C	KELNAKLGNFNNNNNNGLKN**STEPIYA**KVNKKK	22/272
A_D	**KELNE**KLFGNSNNNNNGLKNNT**EPIYA**QVNKKK	53/272
A_D	**XXXXX**KLFGNSNNNNNGLKNNT**EPIYA**QVNKKK	22/272
B_C	**T**GQ**V**ASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG	25/262
B_C	**A**GQ**A**ASPE**EPIYA**QVAKKVNAKIDRLNQIASGLGGVGQAAG	19/262
B_D	TGQ**A**TSPE**EPIYA**QVAKKVSAKIDQLNEATS	25/262
B_D	TGQ**V**ASPE**EPIYA**QVAKKVSAKIDQLNEATS	19/262
C	FPLKRHDKVDDLSKVGR**S**VSP**EPIYA**TIDDLGGP	144/343
C	FPLKRHDKVDDLSKVGR**A**VSP**EPIYA**TIDDLGGP	50/343
D	AINRKIDRINKIASAGKGVGGFSGAGRSASP**EPIYA**TIDFDE**A**N	144/343
D	AINRKIDRINKIASAGKGVGGFSGAGRSASP**EPIYA**TIDFDE**T**N	50/343

[a]X represents unknown amino acids; the amino acids which are different in two sequences shown are highlighted; Ratio = (Frequency of the type)/(Total frequency).
doi:10.1371/journal.pone.0007736.t004

ABDD in a previous study [44]. However, the sequence type was classified as ABB″D in this work. The 3rd segment that differs between the two studies (D vs. B″) is AINRKIDRINKIASA-GKGVGGFSGAGRSASP**EPIYA**QVAKKVSAKIDQLNEATS. In this segment, the part before the EPIYA motif is similar to segment D, whereas the part after the EPIYA motif is similar to segment B. Obviously, this segment is neither D nor B, rather B″, a variant of segment B (Table 2). Overall, we believe that the definitions of segment and the sequence classifications used in this study are more meaningful and accurate than those used in previous studies.

Each of the 560 CagAs was found to have at least one, and as many as seven, EPIYA segments (or EPIYA motifs). The distributions are 3, 27, 416, 86, 23, 3, 2, and 0 for number of sequences containing 1 through 8 EPIYA segments (Table S6), respectively. For example, a sequence of type A has only one EPIYA segment A and a sequence of type ABCCCCC has seven EPIYA segment, including five repeats of segment C. The majority (74% = 416/560) of sequences had three EPIYA segments.

Detailed Analyses of EPIYA Segments

The EPIYA segment types were defined according to the segment patterns (Table 2); however the composite amino acids varied slightly within each segment type. The two most frequent segments in segments A, B, C and D are shown in Table 4. The segments of EPIYA-A_C or -A_D contain from two to eight Ns (Gln) at the upstream of the pseudo EPIYA-A_C or -A_D motif. The segments C and D have higher consensus than segments A_C, A_D, B_C and B_D.

There were obvious differences between segment C and D when analyzed using the program, WebLogo (Figure 2). The segments were aligned using BioEdit before they were entered into WebLogo. As WebLogo had a problem analyzing a column of aligned sequences if BioEdit had added many spaces, all spaces in the sequence alignments were replaced by Z (meaning zero or nothing). In this way, the inserted space (Z) and the minor amino acids were easily identified. In the alignments, X indicates that an amino acid was not-available. As shown in Figure 2, the lengths of segments A_C and A_D are the same and the sequences of segments A_C and A_D are very similar. However the lengths of segments B_C and B_D, and the segments C and D are quite different. The sequences after the stretch of amino acids, QVAKKV, in segments B_C and B_D were highly variable, while the sequences of segments C and D were completely different. Overall, the sequence main variation between Western and East Asian strains starts after QVAKKV in segments B_C and B_D.

The four amino acids TIDD and TIDF following EPIYA motifs in segments C and D are reported to be important for the binding SHP-2 [17,24]; therefore, the frequency of the four amino acids following EPIYA motifs in all EPIYA segments may be useful. As illustrated in Table S7, the sequences, KVNK and QVNK, occupy this position in the majority of segments A_C and A_D, respectively. QVAK occupied this position in most segments B_C and B_D. In the literature [17], the criteria for identifying EPIYA segments C and D are that the EPIYA motif is followed by TIDD and TIDF, respectively. However, by sequence pattern comparison, we found that EPIYA also belongs to segment C if it is followed by TIEE, TIDE, SIDD, TIDG, TIAE, or TIAD. If EPIYA is followed by TIDS, then it belongs to motif type D. As shown in Table S2, the segments B, B′, and B′′ had the biggest changes in their composite five amino acids. However, the four amino acids following the EPIYA motif were most variable in segment A (Table S7).

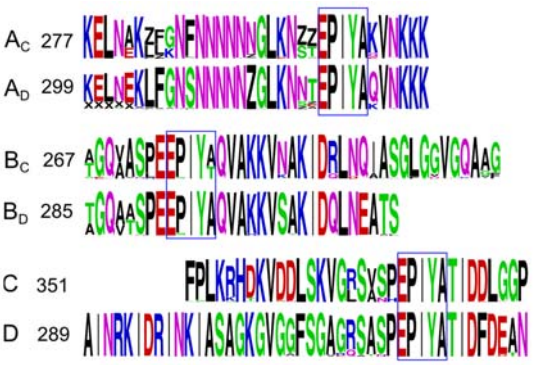

Figure 2. WebLogos of aligned segments of EPIYA-A, -B, and -C/D. The numbers of sequences for each WebLogo are indicated. The sequences were aligned using BioEdit. Z represents space inserted by BioEdit and X represents unknown amino acids.
doi:10.1371/journal.pone.0007736.g002

Correlation of Sequence Types and Geographic Areas

H. pylori strains from different geographic areas are associated with clear phylogeographic differentiation and *H. pylori* populations tend to spread along the lines of human migratory fluxes [45–50]. Furthermore, several studies concluded that CagA isoforms with segments C and D are related to Western and East Asian countries, respectively [14–16]. We tested this hypothesis using our comprehensive system of CagA classification. The frequency of each sequence class in individual countries is shown in Table 5. As expected, all 227 (100%) samples from Western countries contain EPIYA segment C. In contrast, of 307 sequences from East Asian countries (Japan, China, Korean, and Viet Nam), 26 (~8%) contain EPIYA segment C instead of segment D. Interestingly, of the 21 Japanese strains with CagA sequence types related to segment C, 17 have names beginning with OK (Table S8), signifying that they were isolated in Okinawa, Japan (discussed below). The prevalence of sequences containing segments C and D in Southeast Asian countries (Thailand and Malaysia) were the same; and all samples from Iran, Kazakhstan (Kazak), and India were classified as segment C, although they are Asian countries. Overall, we found that it is largely true that CagA with sequences segments C and D are related to Western and East Asian countries, respectively; however, there are some exceptions for East Asian strains. Southeast Asian countries form the geographical border between segment C and segment D. The fact that some East Asian countries have Western type CagA reflects the partial transmission of *H. pylori* from Western to East Asian countries either during the human migration long time ago or recent transmission.

As mentioned above, there are 21 strains from Japan with sequences related to EPIYA segment C instead of segment D (Table 5). The detailed information of these 21 strains is given in Table S8. Most of these segment C strains were isolated from Okinawa, which was governed by the United States from the end of World War II until 1972, and even today there are many US populations living in Okinawa. These data show that transmission of *H. pylori* between different populations may not be a rare event. In fact, previous reports of native Americans in Peru show that all *H. pylori* strains in this population are of the Western type [51], while only 4 of 17 strains isolated from American primitive, an isolated group living in the Amazonian jungles of Colombia, were East Asian type strains [48]. Based on our data, the Western strains are more easily transferred to East Asian people than the other way around. Another possibility for Western type CagA in Okinawa is that the Okinawan CagA is the novel type CagA; the origin did not come from modern Western people, but came to Japan long ago. Further studies will be necessary to test this hypothesis. If it proves true, elucidating the mechanism will be important for understanding the transmission of *H. pylori* in human populations.

Among the 21 strains from Okinawa, 20 contain EPIYA segment B (Table S8). Of 20 EPIYA motifs in segment B, 15 are EPIYT and 4 are ESIYT. Comparing this information with the data in Table S2, we found that the frequencies of the EPIYT and ESIYT motifs among the sequences of the 21 Okinawa strains are also relatively high. Detailed analyses for large number of strains from Okinawa will provide us some information about the roles and evolution of EPIYA motifs.

Table 5. Frequency of CagAs with respect to country[a].

Country	total #	# of seq. containing EPIYA-C	# of seq. containing EPIYA-D
Japan	249	21	228
China	48	4	44
Korea	6	1	5
Viet Nam	4	0	4
Thailand	5	2	3
Malaysia	3	2	1
Iran	5	5	0
India	4	4	0
Kazakhstan	3	3	0
Greece	100	100	0
Italy	34	34	0
Sweden	5	5	0
Ireland	3	3	0
USA	22	22	0
Costa Rica	33	33	0
Colombia	24	24	0

[a]Austria, Chile, and Germany each have one strain. The country information of 11 sequences or strains is not available.

doi:10.1371/journal.pone.0007736.t005

Correlation of Sequence Types and Strain Diseases

We were able to obtain clinical information for 433 strains out of the 560 strains in our data set (Table 6). In our data sheet, disease G contains gastritis, atrophic gastritis, epigastrial pain, gastric hyperplastic polyp, non-ulcer dyspepsia, chronic gastritis, chronic atrophic gastritis, and chronic gastritis-associated dyspepsia as well as "gastris", which are regarded as typo of gastritis. Disease DU and GU (peptic ulcer PU = DU + GU) represent duodenal ulcer and gastric ulcer, respectively. Disease GC contains gastric cancer, gastric carcinoma, gastric adenocarcinoma, gastric adenoma and adenomatous polyps. Disease MALT contains MALT lymphoma and MALToma. Disease E represents esophagitis. Among those 433 samples, 42%, 32%, and 20% of the patients had diseases G, PU, and GC, respectively, which shows that there is a potential for selection bias in the sequence samples. For example, the prevalence of GC is approximately 3% in *H. pylori*-positive patients [52]. Nonetheless, the data are useful when comparing patterns of sequence types among diseases.

We compared three types ABC, ABD and ABCC in relation to clinical outcomes. Other EPIYA types were excluded since the number of other minor types was relatively small. As shown in Table 7, the prevalence of ABCC was 22% (17/[22 + 38 + 17]) in GC; whereas only 12% (18/[65 + 66 + 18]) in G and 7% (8/[42 + 64 + 8]) in PU. The ratio of ABCC/ABC was therefore significantly higher in GC (17/22 = 0.77) than in PU (8/42 = 0.19) and G (18/65 = 0.28) (The calculated chi-square is 8.24 and 6.22, and the probabilities of null hypothesis are less than 0.03 and 0.01, respectively). The data that strains with more

Table 6. Frequency and percentage of strains of certain type disease[a].

Disease	G	DU	GU	GC	E	MALT	Total
Occurrence	181	90	43	87	21	5	433
Percentage	42%	21%	10%	20%	5%	1%	100%

[a]The diseases are designated in the text.

doi:10.1371/journal.pone.0007736.t006

Table 7. EPIYA types and clinical outcomes[a].

	Total	G	PU	GC
ABC	129	65, 50%, 1.0	42, 33%, 1.0	22, 17%, 1.0
ABD	168	66, 39%, 0.8	64, 38%, 1.2	38, 23%, 1.3
ABCC	43	18, 42%, 0.8	8, 19%, 0.6	17, 40%, 2.4

[a]PU = DU + GU. Other diseases are designated in the text. The strains with unavailable disease information are not included.

doi:10.1371/journal.pone.0007736.t007

EPIYA segment C have a greater chance of developing gastric cancer is consistent with previous studies [15,38]. The ratio of ABD/ABC was also higher in GC (38/22 = 1.73) than in PU (64/42 = 1.52) and G (66/65 = 1.02); however the differences were not statistically significant (The calculated chi-square is 0.14 and 2.79, and the probabilities of null hypothesis are more than 0.90 and 0.10, respectively).

The 145, 44, and 169 sequences of types ABC, ABD, and ABCC, respectively, from strains with disease information were used for phylogenic analysis with ClustalW (http://align.genome.jp/). The resulting trees are shown in Table S9, S10 and S11 in the supplementary material. The phylogenetic analysis did not reveal any association between a particular disease and a specific CagA sequence.

Conclusion

In this study, 560 unique CagA sequences containing EPIYA-like motifs were analyzed and in addition to the four previously reported major CagA segment types (A, B, C and D), we found that there are various novel types. Our results allow a clearer classification of the CagA protein sequences and provide a basis for further molecular studies of the pathogenicity of this important protein. In addition, we confirmed that strains with two EPIYA segment C have a greater chance of developing gastric cancer than those with one segment C. However, we did not find any association between a particular disease and specific CagA sequences through phylogenic tree analysis and further studies with larger number of sequences might be necessary whether the specific CagA sequences are involved in the development of clinical outcomes.

Materials and Methods

Data Collection

Three databases, NCBI (National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov), UniProtKB/Swiss-Prot (the Swiss Institute for Bioinformatics and the European Bioinformatics Institute, www.ebi.ac.uk/swissprot/), and DDBJ (DNA Data Bank of Japan, the National Institute of Genetics, www.ddbj.nig.ac. jp/), were used to obtain CagA sequencing data. As of Apr 16, 2007, 1,423 entries were retrieved by searching "protein" at NCBI for "Helicobacter pylori CagA" with display format of "GenPept (Full)". All related data were saved to a local disk. 1,034 entries were retrieved by searching the library, "UniProtKB/Swiss-Prot & UniProtKB/TrEMBL" at Swiss-Prot for "Helicobacter pylori CagA". The related data were downloaded in a "Flat File Format". Similarly, 2,077 entries were retrieved by searching "protein" at DDBJ for "Helicobacter pylori CagA". By choosing "Complete entries", the data were saved as ASCII text on a local disk. The data from DDBJ include the data from NCBI and UniProtKB/Swiss-Prot. We found that the sequences from NCBI included all sequences from UniProtKB/Swiss-Prot and DDBJ; therefore, only the NCBI data were used for sequence analyses. We have collected clinical information for 433 strains related to *H. pylori* CagA. The information is from our data base (from Y.Y.), the NCBI database, and the literature [53,54,18,19].

Data Filtering

EPIYA motifs are located in the C-terminus of the CagA protein. 1,423 entries annotated as CagA in NCBI were downloaded from GenBank. Two rounds of data filtering were used to refine the data obtained from NCBI: (1) removing 832 sequences not containing EPIYA or EPIYA-like motifs (Table S12) and (2) removing 31 redundant sequences (Table S13). Among the 31 sequences, 18 sequences are completely same as others and 13 sequences are parts of others. After the two rounds of filtering, 560 unique CagAs containing EPIYA or EPIYA-like motifs remained (Table S3).

Statistical Analyses

Chi-square test is used to test the statistical significance of the difference of strains of sequence types ABCC and ABC in disease groups GC, PU and G. From Table 7, 17 and 22 strains with ABCC and ABC types appear in disease GC group, and 8 and 42 strains with ABCC and ABC types appear in disease PU group. The calculated chi-square (http://math.hws.edu/javamath/ryan/ChiSquare.html) is 8.24 from a 2×2 matrix. Similarly, 17 and 22 strains with ABCC and ABC types appear in disease GC group, and 18 and 65 strains with ABCC and ABC types appear in disease G group. The calculated chi-square is 6.22 from a 2×2 matrix. Then from a chi-square table, the probabilities of null hypothesis are less than 0.03 and 0.01, respectively, with a df = 1 (df: degree of freedom).

Software for Data Analysis

Home-made program based on MATLAB was used to extract information from the original data retrieved from NCBI, search the sequences, sort the sequences according to disease, create files in FASTA format, etc. BioEdit and WebLogo were used to align and display protein sequences [55,56]. ClustalW (http://align.genome.jp/) and TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview. html) were applied to build and view phylogenic trees.

Supporting Information

Table S1 Full list of representative segments of EPIYA motifs Note: the values in the table are the frequencies of similar sequences, not the number of identical type sequences within a sequence. The highlighted segments are removed in Table 2.
Found at: doi:10.1371/journal.pone.0007736.s001 (0.02 MB XLS)

Table S2 Distribution of EPIYA motifs in segments A, B, C and D
Found at: doi:10.1371/journal.pone.0007736.s002 (0.02 MB XLS)

Table S3 Unique CagA sequences and their sequence types
Found at: doi:10.1371/journal.pone.0007736.s003 (0.10 MB XLS)

Tabld S4 Frequencies of all sequence types All sequence types are listed in Table 3S in supplementary.pdf. The highlighted sequence types are removed in Table 3.

Found at: doi:10.1371/journal.pone.0007736.s004 (0.02 MB XLS)

Table S5 Comparison of sequence classifications in literatures [1] T.Uchida, R. Kanada, Y. Tsukamoto, N. Hijiya, K. Matsuura, S. et al., Cancer Sci. 98 (2007) 521–528. [2] M. Naito, T. Yamazaki, R. Tsutsumi, H. Higashi, K. Onoe, et al., Gastroenterology 130 (2006) 1181–1190.
Found at: doi:10.1371/journal.pone.0007736.s005 (0.02 MB XLS)

Table S6 Distribution of multiple repeats of EPIYA segments
Found at: doi:10.1371/journal.pone.0007736.s006 (0.02 MB XLS)

Table S7 Distribution of the first four amino acids following EPIYA motifs
Found at: doi:10.1371/journal.pone.0007736.s007 (0.02 MB XLS)

Table S8 Samples (from Japan) related to EPIYA-C
Found at: doi:10.1371/journal.pone.0007736.s008 (0.02 MB XLS)

Table S9 The phylogenic tree of fragments ABC
Found at: doi:10.1371/journal.pone.0007736.s009 (0.40 MB XLS)

Table S10 The phylogenic tree of fragments ABCC
Found at: doi:10.1371/journal.pone.0007736.s010 (0.11 MB XLS)

Table S11 The phylogenic tree of fragments ABCC
Found at: doi:10.1371/journal.pone.0007736.s011 (0.40 MB XLS)

Table S12 The information of sequences without EPIYA-like motif
Found at: doi:10.1371/journal.pone.0007736.s012 (0.13 MB XLS)

Table S13 The information of redundant sequences *The sequences under ANo2 are completely same as or cover that under Ano. **Length2 are the length of sequences under ANo2.
Found at: doi:10.1371/journal.pone.0007736.s013 (0.02 MB XLS)

Acknowledgments

Authors thank Dr. Tongbin Li at University of Minnesota for his helpful suggestions in numerous discussions.

Author Contributions

Conceived and designed the experiments: YX YY XG. Performed the experiments: YX YY QZ IM XG. Analyzed the data: YX YY XG. Contributed reagents/materials/analysis tools: YX YY XG. Wrote the paper: YX YY XG.

References

1. Suerbaum S, Michetti P (2002) *Helicobacter pylori* infection. N Engl J Med 347: 1175–1186.
2. Ferreira AC, Isomoto H, Moriyama M, Fujioka T, Machado JC, et al. (2008) *Helicobacter* and gastric malignancies. Helicobacter Suppl 1: 28–34.
3. Franco AT, Johnston E, Krishna U, Yamaoka Y, Israel DA, et al. (2008) Regulation of gastric carcinogenesis by *Helicobacter pylori* virulence factors. Cancer Res 68: 379–87.
4. Hatakeyama M (2008) *Helicobacter pylori* and gastric carcinogenesis. J Gastroenterol 44: 239–48.
5. Rizwan M, Alvi A, Ahmed N (2008) Novel protein antigen (JHP940) from the genomic plasticity region of *Helicobacter pylori* induces tumor necrosis factor alpha and interleukin-8 secretion by human macrophages. J Bacteriol 190: 1146–51.
6. Snider JL, Cardelli JA (2009) *Helicobacter pylori* induces cancer cell motility independent of the c-Met receptor. J Carcinog 8: 7.
7. Umeda M, Murata-Kamiya N, Saito Y, Ohba Y, Takahashi M, et al. (2009) *Helicobacter pylori* CagA causes mitotic impairment and induces chromosomal instability. J Biol Chem. 2009 Jun 22. [Epub ahead of print].
8. Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM, et al. (1995) Infection with *Helicobacter pylori* strains possessing *cagA* is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res 55: 2111–2115.
9. Kuipers EJ, Perez-Perez GI, Meuwissen SGM, Blaser MJ (1995) *Helicobacter pylori* and atrophic gastritis: importance of the *cagA* status. J Natl Cancer Inst 87: 1777–1780.
10. Nomura AM, Lee J, Stemmermann GN, Nomura RY, Perez-Perez GI, et al. (2002) *Helicobacter pylori* CagA seropositivity and gastric carcinoma risk in a Japanese American population. J Infect Dis 186: 1138–1144.
11. Parsonnet J, Friedman GD, Orentreich N, Vogelman H (1997) Risk for gastric cancer in people with CagA positive or CagA negative *Helicobacter pylori* infection. Gut 40: 297–301.
12. Covacci A, Censini S, Bugnoli M, Petracca R, Burroni D, et al. (1993) Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. Proc Natl Acad Sci USA 90: 5791–5795.
13. Tummuru MK, Cover TL, Blaser MJ (1993) Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori:* evidence of linkage to cytotoxin production. Infect Immun 61: 1799–1809.
14. Yamaoka Y, Kodama T, Kashima K, Graham DY, Sepulveda AR (1998) Variants of the 3′ region of the *cagA* gene in *Helicobacter pylori* isolates from patients with different *H. pylori*-associated diseases. J Clin Microbiol 36: 2258–2263.
15. Yamaoka Y, El-Zimaity HM, Gutierrez O, Figura N, Kim JG, et al. (1999) Relationship between the *cagA* 3′ repeat region of *Helicobacter pylori*, gastric histology, and susceptibility to low pH. Gastroenterology 117: 342–349.
16. Yamaoka Y, Osato MS, Sepulveda AR, Gutierrez O, Figura N, et al. (2000) Molecular epidemiology of *Helicobacter pylori*: separation of *H. pylori* from East Asian and non-Asian countries. Epidemiol Infect 124: 91–96.
17. Higashi H, Tsutsumi R, Fujita A, Yamazaki S, Asaka M, et al. (2002) Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. Proc Natl Acad Sci USA 99: 14428–14433.
18. Satomi S, Yamakawa A, Matsunaga S, Masaki R, Inagaki T, et al. (2006) Relationship between the diversity of the *cagA* gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan. J Gastroenterol 41: 668–673.
19. Yamazaki S, Yamakawa A, Okuda T, Ohtani M, Suto H, et al. (2005) Distinct diversity of *vacA*, *cagA*, and *cagE* genes of *Helicobacter pylori* associated with peptic ulcer in Japan. J Clin Microbiol 43: 3906–3916.
20. Hatakeyama M (2004) Oncogenic mechanisms of the *Helicobacter pylori* CagA protein. Nat Rev Cancer 4: 688–694.
21. Higashi H, Yokoyama K, Fujii Y, Ren S, Yuasa H, et al. (2005) EPIYA motif is a membrane-targeting signal of *Helicobacter pylori* virulence factor CagA in mammalian cells. J Biol Chem 280: 23130–23137.
22. Hatakeyama M (2006) *Helicobacter pylori* CagA – a bacterial intruder conspiring gastric carcinogenesis. Int J Cancer 119: 1217–1223.
23. Naito M, Yamazaki T, Tsutsumi R, Higashi H, Onoe K, et al. (2006) Influence of EPIYA-repeat polymorphism on the phosphorylation-dependent biological activity of *Helicobacter pylori* CagA. Gastroenterology 130: 1181–1190.
24. Backert S, Moese S, Selbach M, Brinkmann V, Meyer TF (2001) Phosphorylation of tyrosine 972 of the *Helicobacter pylori* CagA protein is essential for induction of a scattering phenotype in gastric epithelial cells. Mol. Microbiol. 42: 631–644.
25. Censini S, Lange C, Xiang Z, Crabtree JE, Ghiara P, et al. (1996) *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. Proc Natl Acad Sci USA 93: 14648–14653.
26. Asahi M, Azuma T, Ito S, Ito Y, Suto H, et al. (2000) *Helicobacter pylori* CagA protein can be tyrosine phosphorylated in gastric epithelial cells. J Exp Med 191: 593–602.
27. Backert S, Ziska E, Brinkmann V, Zimny-Arndt U, Fauconnier A, et al. (2000) Translocation of the *Helicobacter pylori* CagA protein in gastric epithelial cells by a type IV secretion apparatus. Cell Microbiol 2: 155–164.
28. Odenbreit S, Püls J, Sedlmaier B, Gerland E, Fischer W, Haas R al (2000) Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. Science 287: 1497–1500.
29. Segal ED, Cha J, Lo J, Falkow S, Tompkins LS, et al. (1999) Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by *Helicobacter pylori*. Proc Natl Acad Sci USA 96: 14559–14564.
30. Stein M, Rappuoli R, Covacci A (2000) Tyrosine phosphorylation of the *Helicobacter pylori* CagA antigen after cag-driven host cell translocation. Proc Natl Acad Sci USA 97: 1263–1268.
31. Higashi H, Tsutsumi R, Muto S, Sugiyama T, Azuma T, et al. (2002) SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein. Science 295: 683–686.
32. Backert S, Moese S, Selbach M, Brinkmann V, Meyer TF (2001) Phosphorylation of tyrosine 972 of the *Helicobacter pylori* CagA protein is essential for

induction of a scattering phenotype in gastric epithelial cells. Mol Microbiol 42: 631–644.
33. Tammer I, Brandt S, Hartig R, König K, Backert S (2007) Activation of Abl by *Helicobacter pylori*: A Novel kinase for CagA and crucial mediator of host cell scattering. Gastroenterology 132: 1309–1319.
34. Selbach M, Paul FE, Brandt S, Guye P, Daumke O, et al. (2009) Host cell interactome of tyrosine-phosphorylated bacterial proteins. Cell Host & Microbe 5: 397–403.
35. Judd LM, Alderman BM, Howlett M, Shulkes A, Dow C, et al. (2004) Gastric cancer development in mice lacking the SHP2 binding site on the IL-6 family co-receptor gp130. Gastroenterology 126: 196–207.
36. Tebbutt NC, Giraud AS, Inglese M, Jenkins B, Waring P, et al. (2002) Reciprocal regulation of gastrointestinal homeostasis by SHP2 and STAT-mediated trefoil gene activation in gp130 mutant mice. Nat Med 8: 1089–1097.
37. Argent RH, Hale JL, EL-Omar EM, Atherton JC, et al. (2008) Differences in *helicobacter pylori cagA* tyrosine phosphorylation motif patterns between western and east asian strains, and influences on interleukin-8 secretion. J med microbiol 57: 1062–1067.
38. Argent RH, Kidd M, Owen RJ, Thomas RJ, Limb MC, et al. (2004) Determinants and consequences of different levels of CagA phosphorylation for clinical isolates of *Helicobacter pylori*. Gastroenterology 127: 514–523.
39. Azuma T, Yamakawa A, Yamazaki S, Fukuta K, Ohtani M, et al. (2002) Correlation between variation of the 3′ region of the *cagA* gene in *Helicobacter pylori* and disease outcome in Japan. J Infect Dis 186: 1621–1630.
40. Kidd M, Lastovica AJ, Atherton JC, Louw JA, et al. (1999) Heterogeneity in the *Helicobacter pylori vacA* and *cagA* genes: association with gastroduodenal disease in South Africa? Gut 45: 499–502.
41. Rota CA, Pereira-Lima JC, Blaya C, Nardi NB (2001) Consensus and variable region PCR analysis of *Helicobacter pylori* 3′ region of *cagA* gene in isolates from individuals with or without peptic ulcer. J Clin Microbiol 39: 606–612.
42. Jenks PJ, Mégraud F, Labigne A (1998) Clinical outcome after infection with *Helicobacter pylori* does not appear to be reliably predicted by the presence of any of the genes of the *cag* pathogenicity island. Gut 43: 752–758.
43. Zhou J, Zhang J, Xu C, He L (2004) *cagA* genotype and variants in Chinese *Helicobacter pylori* strains and relationship to gastroduodenal diseases. J Med Microbiol 53: 231–235.
44. Uchida T, Kanada R, Tsukamoto Y, Hijiya N, Matsuura K, et al. (2007) Immunohistochemical diagnosis of the *cagA*-gene genotype of *Helicobacter pylori* with anti-East Asian CagA-specific antibody. Cancer Sci 98: 521–528.
45. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, et al. (2003) Traces of human migrations in *Helicobacter pylori* populations. Science 299: 1582–1585.
46. Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, et al. (2005) Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. PLoS Genet 1: e43.
47. Linz B, Balloux F, Moodley Y, Manica A, Liu H, et al. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. Nature 445: 915–918.
48. Yamaoka Y, Orito E, Mizokami M, Gutierrez O, Saitou N, et al. (2002) *Helicobacter pylori* in North and South America before Columbus. FEBS Lett 517: 180–184.
49. Yamaoka Y, Kato M, Asaka M (2008) Geographic differences in gastric cancer incidence can be explained by differences between *Helicobacter pylori* strains. Intern Med 47: 1077–1083.
50. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, et al. (2009) The peopling of the pacific from a bacterial perspective. Science 323: 527–530.
51. Kersulyte D, Mukhopadhyay AK, Velapatiño B, Su W, Pan Z, et al. (2000) Differences in genotypes of *Helicobacter pylori* from different human populations. J Bacteriol 182: 3210–3218.
52. Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, et al. (2001) *Helicobacter pylori* infection and the development of gastric cancer. N Engl J Med 345: 784–789.
53. Azuma T, Yamakawa A, Yamazaki S, Ohtani M, Ito Y, et al. (2004) Distinct diversity of the *cag* pathogenicity island among *Helicobacter pylori* strains in Japan. J Clin Microbiol 42: 2508–2517.
54. Hoshino FB, Katayama K, Watanabe K, Takahashi S, Uchimura H, et al. (2000) Heterogeneity found in the *cagA* gene of *Helicobacter pylori* from Japanese and non-Japanese isolates. J Gastroenterol 35: 890–897.
55. Hall TA (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp 41: 95–98.
56. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.