Kobe J. Med. Sci., Vol. 66, No. 1, pp. E22-E31, 2020

Detection of Novel Amino Acid Polymorphisms in the East Asian CagA of *Helicobacter Pylori* with Full Sequencing Data

HIROKI HAYASHI[1], JUN INOUE[1]*, KATSUAKI OYAMA[1], KOKI MATSUOKA[1], SHIN NISHIUMI[1], MASARU YOSHIDA[1,2], YOSHIHIKO YANO[1,3], and YUZO KODAMA[1]

[1]*Division of Gastroenterology, Department of Internal Medicine, Kobe University Graduate School of Medicine, Kobe, Japan. *Corresponding author;*

[2]*Division of Metabolomics Research, Department of Internal Related, Kobe University Graduate School of Medicine, Kobe, Japan;*

[3]*Division of Molecular Medicine & Medical Genetics, Department of Pathology, Kobe University Graduate School of Medicine, Kobe, Japan.*

Received 5 February 2020/ Accepted 17 March 2020

Key words: *Helicobacter pylori*, East Asian CagA, Amino Acid Polymorphisms, N-terminal Region, *In Silico*

Cytotoxin-associated gene A (CagA) is generally accepted to be the most important virulence factor of *Helicobacter pylori* and increases the risk of developing gastric cancer. East Asian CagA, which includes the EPIYA-D segment at the C-terminal region, has a significantly higher gastric carcinogenic rate than Western CagA including the EPIYA-C segment. Although the amino acid polymorphism surrounding the EPIYA motif in the C-terminal region has been examined in detail, limited information is currently available on the amino acid polymorphism of the N-terminal region of East Asian CagA. In the present study, we analyzed the sequencing data of East Asian CagA that we obtained previously to detect amino acid changes (AACs) in the N-terminal region of East Asian CagA. Four highly frequent AACs in the N-terminal region of East Asian CagA were detected in our datasets, two of which (V356A, Y677F) exhibited reproducible specificity using a validation dataset from the NCBI database, which are candidate AACs related to the pathogenic function of CagA. We examined whether these AACs affect the functions of CagA *in silico* model. The computational docking simulation model showed that binding affinity between CagA and phosphatidylserine remained unchanged in the model of mutant CagA reflecting both AAC, whereas that between CagA and α5β1 integrin significantly increased. Based on whole genome sequencing data we herein identified novel specific AACs in the N-terminal regions of EPIYA-D that have the potential to change the function of CagA.

INTRODUCTION

Helicobacter pylori (*H. pylori*) is a Gram-negative microaerophilic bacterium that has been implicated in the etiology of gastric carcinogenesis in several epidemiological and molecular studies (32). Chronic infection of the gastric mucosa by *H. pylori* is widely accepted to cause various diseases, including atrophic gastritis, gastroduodenal ulcers, gastric B-cell lymphomas, and gastric cancer (27, 38, 33).

Cytotoxin-associated gene A (CagA) is generally considered to be the most important virulence factor of *H. pylori*, and its expression is closely associated with an increased risk of developing atrophic gastritis and gastric cancer (7, 28). *CagA* is located in the cag pathogenicity island (PAI) region of the *H. pylori* genome (2). CagA is translocated into gastric cells through the type IV secretion system (T4SS) or an endocytic pathway and plays an important role in the development of gastric cancer (10, 9, 34). The CagA protein consists of approximately 1,200 amino acids divided into the N- and C-terminal regions (40). After translocating into host gastric cells, the CagA protein localizes to the inner leaflet of the plasma membrane and undergoes tyrosine phosphorylation by Src and Abl family kinases (5). Tyrosine-phosphorylated CagA binds to and activates Src homology 2 domain-containing protein tyrosine phosphatase 2 (SHP2) in host gastric cells (16). Activated SHP2 modifies the signal pathways that induce abnormal cell proliferation and elongation, which are essential processes in the development of gastric cancer (30).

Tyrosine phosphorylation targets are Glu-Pro-Ile-Tyr-Ala (EPIYA) motifs located at the C-terminal region of CagA (17). EPIYA segments are classified as EPIYA-A, -B, -C, or -D based on the amino acid sequence surrounding each of these motifs (13). The *cagA* gene is mainly grouped into two different allele types: the East Asian type that contains the EPIYA-D segment after the EPIYA-A and EPIYA-B segments and the Western type that contains one to three repeats of the EPIYA-C segment after the EPIYA-A and EPIYA-B segments (16, 15,

Phone: +81-78-382-6305 Fax: +81-78-382-6309 E-mail: juninoue@med.kobe-u.ac.jp

23). East Asian type CagA is associated with a higher risk of gastric cancer than the Western type (4, 3, 31). One of the factors contributing to the high virulence of East Asian CagA is stronger SHP-2-binding activity resulting from the different amino acid sequence of the EPIYA type; therefore, many studies have focused on nucleotide polymorphisms surrounding the EPIYA motif at the C-terminal region. However, recent structural and functional analyses indicated that the N-terminal region of CagA contains three domains: Domains I, II, and III. Domain I (aa 24-221) is the extreme N-terminal region that consists of 10 α helices and is structurally isolated from the other two domains (15, 14). Domain I contributes to gastric carcinogenesis by inhibiting tumor suppressor proteins in host cells, such as ASSP2 and RUNX3 (8, 35). Domain II (aa 303-644) has a large antiparallel β sheet and two subdomains of α helices. In helix α18, the residues K613, K614, K617, K621, R624, R626, K631, K635, and K636 constitute a basic amino acid cluster that provides a positive electrostatic surface potential and interacts with negatively charged phosphatidylserine (PS) in the inner leaflet of the plasma membrane (15). This electrostatic interaction influences the strength of CagA binding with PS and is important for not only the delivery of CagA into gastric epithelial cells, but also its localization to the inner leaflet of the plasma membrane (24). Domain II also has a β sheet that binds with the ectodomain of α5β1 integrin in host cells (20, 19). CagA itself has an ability of attaching to gastric epithelial cells by binding to β1 integrin and PS on the outer membrane of host cell and triggers its own uptake into gastric epithelial cells without infection of *H. pylori* (34). Domain III (aa 645-824) is composed of 4 α helices (15). In Domain III, an N-terminal binding sequence (NBS) bundles with the C-terminal binding sequence (CBS) located in the disordered C-terminal tail, creating a C-terminal lariat loop that strengthens the interaction between CagA and SHP2 (14). Despite the N-terminal region of CagA playing its own role in the development of gastric cancer, AACs at the N-terminal region of CagA related to the high virulence of East Asian *H. pylori* remain unclear.

The aim of the present study was to detect AACs at the N-terminal region of East Asian CagA using our previous whole-genome sequencing (WGS) data and a dataset from the National Center for Biotechnology Information (NCBI) database as candidate AACs related to the pathogenic function of East Asian CagA.

MATERIALS AND METHODS

Data collection and EPIYA segment type classification

All data used in this study were obtained from public databases and cannot be linked to patient information. Previous fastq data on 43 *H. pylori* isolates obtained by WGS using Miseq (Illumina) were downloaded from the DNA Data Bank of Japan (DDBJ) under accession numbers DRA001250, DRA002946, and DRA004713. We used CLC Genomics Workbench 8.5.3 (CLC bio, Aarhus, Denmark) to assemble sequence reads. Sequence reads were mapped to ATCC26695 (NC_000915) and F30 (AP011941), which are representative strains of Western strains (EPIYA ABC) and East Asian strains (EPIYA ABD), respectively, and EPIYA segment types were then classified by the amino acid sequence surrounding the EPIYA motif (13). *H. pylori* strains were divided into two groups: the East Asia group and another group, depending on whether the *H. pylori* strain had the EPIYA D segment.

Detection of single nucleotide variants (SNVs)

In order to detect SNVs, sequence reads were mapped against the reference genome of ATCC 26695 (NC_000915), a representative of Western *H. pylori* strains, and SNVs were identified using Fixed Ploidy Variant Detection modules with default parameters and minor modifications to the mapping algorithm. To exclude false-positive variants due to sequencing errors, we selected variants present in > 90.0% of mapped reads. Specific SNVs in the East Asia group were detected with Fisher's exact test, setting the threshold for a *P* value of 0.01. Insertions, deletions, and successive multi-nucleotide variants were excluded as described previously (18).

Prediction of the CagA-V356A-Y677F conformation

The nucleotide sequence of CagA in ATCC26695, a representative of Western *H. pylori* strains, was retrieved with the NCBI accession number NC_000915 and was then translated to amino acids using CLC Genomics Workbench v 8.5.3 (CLC bio). To obtain the predicted conformation of mutant CagA that reflects specific AACs detected in the N-terminal region of East Asian CagA, homology modeling employing SWISS-MODEL (6) was performed using the crystal model of CagA (4DVY: PDB-ID of CagA in ATCC26695, a representative of Western *H. pylori* strains), in which valine at 356 and tyrosine at 677 were substituted into alanine and phenylalanine, respectively. The minimization and removal of the disordered region were then implemented to CagA-V356A-Y677F from SWISS-MODEL by Swiss-PdbViewer v 4.1.0 (12) and PyMOL Molecular Graphics System v 2.2 (Schrödinger, LLC).

Docking simulation of CagA with PS and α5β1 integrin

After detecting specific AACs in East Asian CagA computational docking simulations were performed (Figure 1). The crystal structure of CagA with ATTC26695, a representative of Western *H. pylori* strains, was downloaded from the Protein Data Bank (PDB-ID: 4DVY). SwissDock was used to dock CagA with PS, as described by Ulloa-Guerrero *et al.*, 2018 (36, 11). The files for crystal CagA and CagA-V356A-Y677F were uploaded to the SwissDock web server. To observe the interaction between helix α18 and PS, the web server was set to "Accurate" and run with the following parameters: X center: -32.746, Y center: 2.921, Z center: -21.443, X size: 12, Y size: 22, Z size: 24. The models interacting with helix α18 and PS were selected using Chimera (29), and the difference between the crystal CagA and CagA-V356A-Y677F was assessed by ΔGs from SwissDock using the Wilcoxon test.

The open form of α5β1 integrin, which is a functionally active conformational state (25), was obtained from PDB (PDB-ID: 3VI4). The pdb files of CagA, CagA-V356A-F677F, and α5β1 integrin were subjected to the ClusPro 2.0 web server (21, 22, 37) setting the residues CagA (β-sheet): 304-544 aa, except for 368-446 aa (15), α5 integrin (β-propeller): 1-450 aa (25), and β1 integrin (βA domain): 123-361 aa (25) as binding sites. The 10 best docked models of Balance were then downloaded and potential energy was calculated through Swiss-PdbViewer v 4.1.0 with the Gromos96 force-field *in vacuo* (12). The total energy of the docked models was given by $E_{total} = E_{bound} + E_{angle} + E_{torsion} + E_{improper} + E_{non\text{-}bonded} + E_{elec,}$ where E_{bound}, E_{angle}, $E_{torsion}$, $E_{improper}$, $E_{non\text{-}bonded}$, and E_{elec} are bound, angle, torsion, improper, non-bonded, and electrostatic energies, respectively.

The energy of each docked model was analyzed by the Wilcoxon test. Models including steric hindrance detected by Swiss-PdbViewer were removed from the analysis.

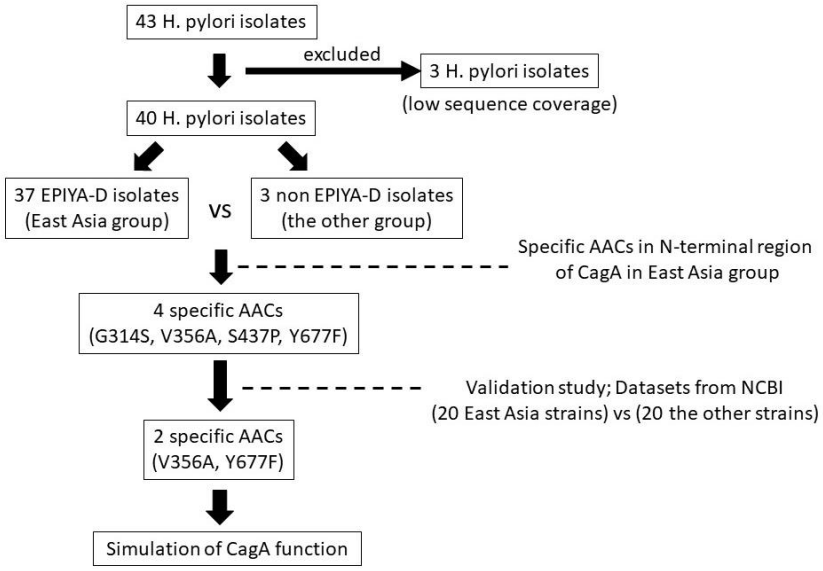

Figure 1. Flowchart of AACs detection in N-terminal region of East Asian CagA
The flowchart shows the process of detecting AACs in N-terminal region of East Asian CagA.

Availability of data and material

Sequence data that support the present results have been deposited in the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/index-e.html) with the accession numbers DRA001250, DRA002946, and DRA004713. Sequence data used as the validation dataset in the present study have been deposited in the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/). The crystal structure of the N-terminal region of CagA has been deposited in the Protein data bank (http://www.rcsb.org/) with PDB-ID: 4DVY. The datasets analyzed during this study are available in a previous study (DOI 10.1186/s13099-017-0165-1) (18).

Statistical analysis

The significance of differences of each SNV was tested using Fisher's exact test, while that of each AAC in the validation set was tested using Pearson's chi-squared test with Bonferroni corrections. ΔG and the total energy of each protein complex were compared by the Wilcoxon test using JMP 10 (SAS). A difference associated with a P value < 0.01 was considered to be significant.

RESULTS

Quality check and mapping of sequence reads in the *cagA* region

Previous sequencing data for 43 clinical *H. pylori* isolates in our past papers are publicly available from the DDBJ database (18, 26). We downloaded the total reads of all 43 *H. pylori* isolates from DDBJ and checked the quality and quantity of sequencing reads in each *H. pylori* isolate to confirm whether these sequencing reads may be used for the identification of genetic variants. Following quality filtering and data trimming, the total reads of each isolate ranged between 1.52 and 9.80 million. Sequencing reads were mapped to the genome of *H. pylori* strain ATCC 26695 (NC_000915), a representative of Western strains, as a reference. The average coverage of high quality reads in the total consensus ranged between 87- and 491-fold (Table I). A focus on read coverage in the *cagA* region of each isolate revealed that strains 189, 194, and S23 had lower coverage (average of less than 70-fold) in the *cagA* region (8.5-, 1.3-, and 65.8-fold, respectively), which was insufficient for the identification of genetic variants in the bacterial genome (18), and, thus, we excluded the data on these three isolates. The average coverage of the remaining 40 isolates ranged between 77- and 341.5-fold (more than 70-fold) for *cagA*, which allowed the detection of SNVs (18). The remaining 40 *H. pylori* isolates were analyzed in more detail. Among the 40 isolates, 28, 7, and 5 were derived from Japanese, Chinese, and Vietnamese patients, respectively.

Table I. Characteristics of 43 *H. pylori* isolates and sequence reads

Strain ID	Region	Total reads	Average coverage (fold)	*cagA* (fold)
F21	Fukui	3,464,871	210	158.2
F23	Fukui	5,294,208	287	222.1
F24	Fukui	4,285,830	249	145.3
F28	Fukui	5,219,931	266	226.4
F32	Fukui	4,450,660	290	265.8
F44	Fukui	2,760,331	148	113.1
F51	Fukui	3,734,580	211	209
F52	Fukui	4,042,321	240	165.6
F57	Fukui	5,778,614	289	242.1
F65	Fukui	3,174,372	180	134
F75	Fukui	4,657,569	237	174.9
F79	Fukui	2,776,184	138	123.3
F94	Fukui	3,001,825	165	137.3
F214	Fukui	4,879,043	245	195
F215	Fukui	4,087,432	213	171.4
F229	Fukui	3,804,388	209	155.1
S1	Kobe	2,001,937	110	93.7
S2	Kobe	4,863,889	245	180.4
S4	Kobe	3,629,063	193	144.1
S8	Kobe	9,798,482	491	341.5
S13	Kobe	3,603,220	190	140.2
S16	Kobe	6,273,095	323	225.2
S17	Kobe	4,543,326	239	205.2
S22	Kobe	5,977,483	305	209.3
S23	Kobe	1,668,471	92	65.8
S26	Kobe	6,194,755	319	222.5
174	Okinawa	2,630,310	111	105.4
177	Okinawa	1,919,866	111	77
179	Okinawa	2,074,398	137	131.8
189	Okinawa	1,518,202	87	8.5
194	Okinawa	6,275,364	289	1.3
HZ2	Hang Zhou	3,695,218	192	151.2
HZ11	Hang Zhou	5,559,310	319	262.5
HZ21	Hang Zhou	5,423,070	300	263.5
HZ34	Hang Zhou	5,238,837	271	237.6
HZ53	Hang Zhou	5,573,514	292	316.6
HZ67	Hang Zhou	3,141,035	193	126.3
HZ82	Hang Zhou	3,630,153	210	163
VN8	Ho Chi Minh	5,393,465	302	150.2
VN17	Ho Chi Minh	6,228,044	349	266.9
VN19	Ho Chi Minh	3,741,401	215	170.2
VN24	Ho Chi Minh	3,863,748	223	170.9
VN27	Ho Chi Minh	3,779,091	211	163.9

The sequence reads of 43 *H. pylori* isolates previously obtained from Fukui, Kobe, Okinawa, Hang Zhou, and Ho Chi Minh were downloaded from DDBJ and total reads after trimming were mapped to ATCC26695 (NC_000915) as a reference and calculated with Genomic Workbench 8.5.3. After mapping, the average coverage of high quality reads in the total consensus and *cagA* region were calculated.

EPIYA segment type classification

To divide 40 isolates into two groups, East Asia group and the other group, *H. pylori* isolates were classified by the EPIYA segment type obtained using the amino acid sequence surrounding the EPIYA motif (13). Sequence reads of 40 *H. pylori* isolates were mapped to the reference strains of the Western and East Asian strains. Among the 40 *H. pylori* isolates, the sequence type of 37 *H. pylori* isolates was East Asian CagA including EPIYA-A, B and D, while that of two *H. pylori* isolates (F79 and HZ53) was Western CagA with EPIYA-A, B and C. However, the remaining *H. pylori* isolate (F65) was not classified as ABD or ABC. By referring to a previous study using Sanger sequencing (3), the *cagA* sequence type of F65 was ABB (Table II). Forty *H. pylori* strains were divided into two groups: the East Asia group including 37 strains and the other group including 3 strains (F65, F79, and HZ53) (Figure 2a).

Table II. Classification of the EPIYA sequence type of 40 isolates

Strain ID	Seq. Type	Strain ID	Seq. Type	Strain ID	Seq. Type
174	ABD	F24	ABD	HZ2	ABD
177	ABD	F28	ABD	HZ11	ABD
179	ABD	F32	ABD	HZ21	ABD
S1	ABD	F44	ABD	HZ34	ABD
S2	ABD	F51	ABD	HZ53	ABC
S4	ABD	F52	ABD	HZ67	ABD
S8	ABD	F57	ABD	HZ82	ABD
S13	ABD	F65	ABB	VN8	ABD
S16	ABD	F75	ABD	VN17	ABD
S17	ABD	F79	ABC	VN19	ABD
S22	ABD	F94	ABD	VN24	ABD
S26	ABD	F214	ABD	VN27	ABD
F21	ABD	F215	ABD		
F23	ABD	F229	ABD		

EPIYA types were obtained based on the amino acid sequence surrounding the EPIYA motif using ATCC26695 (NC_000915), F30 (AP011941) as a reference. The EPIYA sequence type of F65 was confirmed to be ABB using our previous Sanger sequencing data.

Specific AACs in the N-terminal region of East Asian CagA

To analyze AACs in the N-terminal region of CagA common to the East Asia group including 37 strains, consensus amino acid sequences were compared between the East Asia group and the other group. After mapping the sequence reads of each isolate to the ATCC 26695 genome (NC_000915), a representative of Western *H. pylori* strains, as a reference, the SNVs in the N-terminal region of CagA of each strain against the reference sequence were detected and specific SNVs in the East Asia group were identified using Fisher's exact test. Seven highly frequent SNVs were detected in the East Asia group, four of which produced AACs (G314S, V356A, S437P, and Y677F) (Table III).

In order to confirm the specificity of these four AACs of N-terminal region of CagA in the East Asia group, we verified reproducibility using a validation dataset. The amino acid sequence data of CagA in 20 East Asian type *H. pylori* strains and 20 strains with the other type were obtained from NCBI in turn from the top using the criteria "*Helicobacter pylori* CagA complete CDS" in the nucleotide database (Table IV). By using these strains as the validation set, two out of four AACs (V356A and Y677F) were detected more significantly in the East Asia group (Table V). A focus on the positions of these 2 AACs in the ATCC26695 *cagA* sequences showed that V356A was located in Domain II of N-terminal region and Y677F in Domain III of N-terminal region (Figure 2b,c).

Table III. Highly frequent AACs of N-terminal region of CagA in the East Asia group

Position	Reference	Variant	East Asia group		The other group		p value
			(n=37)	(%)	(n=3)	(%)	
314	G	S	37/37	100	1/3	33.3	<0.01
356	V	A	36/37	97.3	0/3	0	<0.01
437	S	P	35/37	94.6	0/3	0	<0.01
677	Y	F	32/37	86.5	0/3	0	<0.01

Using the dataset we previously sequenced by WGS and deposited in DDBJ, the frequency of AACs of N-terminal region of CagA in the East Asia group, including 37 strains, and in the other group, including 3 strains, were compared. Four specific AACs of N-terminal region of CagA were detected in the East Asia group. A statistical analysis was performed using Fisher's exact test. *P*<0.01 indicates a significant difference.

Table IV. The list of *H. pylori* strain ID used as validation dataset

Strain ID							
East Asia group				The other group			
CPY2052	GZ25	J-187	J-230	ATCC43526	Ca73	GZ9	NCTC11639
GZ20	GZ26	J-194	J-241	ATCC43579	Du23:2	GZ10	OK107
GZ21	GZ27	J-198	J-248	ATCC49503	Du52:2	GZ14	OK112
GZ23	J-149	J-207	J-566	B147	F79	GZ22	PMSS1
GZ24	J-16	J-216	J-578	Ca52	GZ5	NCTC11637	SS1

The amino acid sequence data of CagA in 20 East Asian type *H. pylori* strains and 20 strains with the other type were obtained from NCBI in turn from the top using the criteria "*Helicobacter pylori* CagA complete CDS".

Table V. Highly frequent AACs of N-terminal region of CagA in the East Asia group in the validation dataset

Position	Reference	Variant	East Asia group		the other group		p value
			(n=20)	(%)	(n=20)	(%)	
314	G	S	4/20	20	0/20	0	NS
356	V	A	18/20	90	0/20	0	<0.01
437	S	P	1/20	5	0/20	0	NS
677	Y	F	20/20	100	0/20	0	<0.01

Using the dataset obtained from NCBI including 20 strains (East Asia group) and 20 strains (the other group), the reproducibility of these four AACs in the N-terminal region of CagA was verified. The frequency of detection of two out of the four AACs (V356A and Y677F) was significantly higher in the East Asia group. A statistical analysis was performed using Pearson's chi-squared test with Bonferroni corrections. *P*<0.01 indicates a significant difference. NS: not significant.

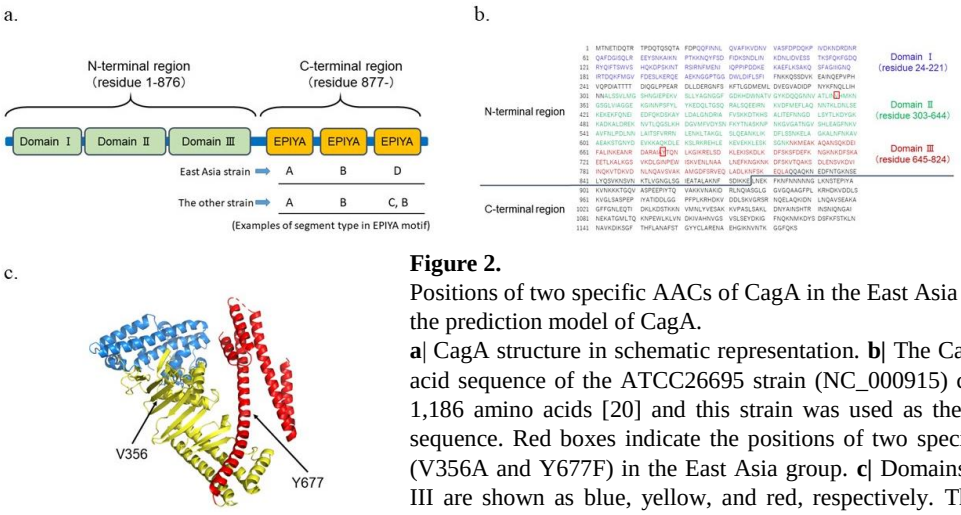

Figure 2.
Positions of two specific AACs of CagA in the East Asia group and the prediction model of CagA.
a| CagA structure in schematic representation. **b**| The CagA amino acid sequence of the ATCC26695 strain (NC_000915) consists of 1,186 amino acids [20] and this strain was used as the reference sequence. Red boxes indicate the positions of two specific AACs (V356A and Y677F) in the East Asia group. **c**| Domains I, II, and III are shown as blue, yellow, and red, respectively. The tertiary structure of the N-terminal region of the reference CagA (PDB: ID 4DVY) was described by PyMol. Arrows indicate the positions of two specific AACs in the East Asia group.

Docking simulation of mutant CagA with PS and α5β1 integrin

To reflect these specific AACs in the N-terminal region of East Asian CagA to the tertiary structure for investigating mutation effects, the CagA mutant, which was named CagA-V356A-Y677F, was made by the SWISS-MODEL server from the crystal model CagA (4DVY: PDB-ID of CagA in ATCC26695, a representative of Western *H. pylori* strains) with the substitution of valine at 356 and tyrosine at 677 into alanine and phenylalanine, respectively. The root mean square deviation (RMSD) of CagA with CagA-V356A-Y677F was 0.267 Å. The binding affinity of helix α18 of CagA with PS in the inner leaflet of host gastric cells is considered to be important for the pathogenicity of CagA (24). To examine the interacting force between PS and helix α18 of CagA, the docking simulation by SwissDock was performed using the crystal model CagA (PDB-ID:4DVY) as a control and CagA-V356A-Y677F as the mutant model. ΔG (kcal/mol) of the CagA docking models with PS were calculated. No significant differences were observed between the crystal and CagA-V356A-Y677F (Figure 3a).

The β sheet of CagA binds with the α5β1 integrin of host cells and this binding is not necessary, but is important and supportive for the process of CagA translocation into host cells (19, 34). To investigate mutational effects on CagA-V356A-Y677F, docking simulations with α5β1 integrin were performed using the ClusPro server. The total energy of the top 10 models from ClusPro was evaluated by Swiss-PdbViewer. The energy of each model was analyzed by the Wilcoxon test (Figure 3b). The energy of CagA-V356A-Y677F was significantly lower than that of the crystal model of CagA (4DVY as a control). This result indicates that the complex of CagA-V356A-F677F with α5β1 integrin is more stable. Therefore, the two detected AACs of N-terminal region of CagA in the East Asia Group may be polymorphisms that increase the binding affinity of CagA to α5β1 integrin.

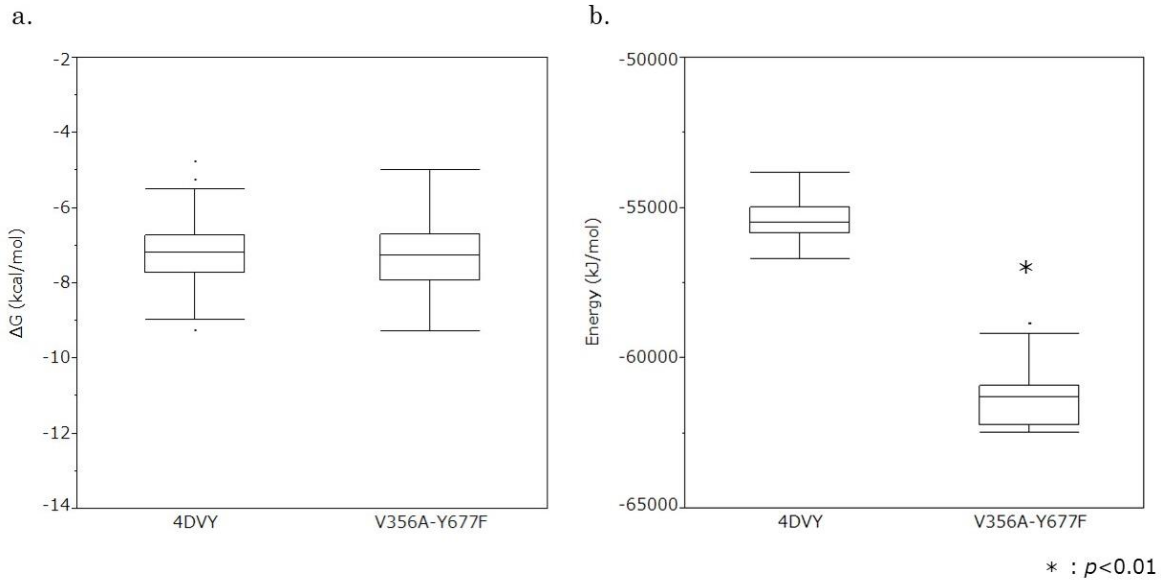

Figure 3. A docking simulation of CagA with PS or α5β1 integrin.
a| ΔG of the docked model with the SwissDock between PS and helix α18 of CagA in each CagA model, the crystal model of CagA (4DVY: PDB-ID of CagA in ATCC26695, a representative strain of Western *H. pylori*), and the mutant model of CagA (CagA-V356A-Y677F). Box plots indicate the 25th, 50th, and 75th percentiles. **b|** Total energy of the docked model with ClusPro 2.0 web server between α5β1 integrin and the β sheet of CagA in each CagA model, the crystal model of CagA (4DVY), and the mutant model of CagA (CagA-V356A-Y677F), were calculated through Swiss-Pdb viewer. The average energy of the docked model between α5β1 integrin and crystal CagA (4DVY) is -55269±801 (kJ/mol), while that of the docked model between α5β1 integrin and mutant CagA (CagA-V356A-Y677F) is -61084±1236 (kJ/mol). Significantly lower energy was obtained from CagA-V356A-Y677F (P<0.01). Box plots indicate the 25th, 50th, and 75th percentiles.

DISCUSSION

H. pylori CagA exhibits oncoprotein activity in mammals and plays a major role in the pathogenesis of *H. pylori* infections in humans. East Asian CagA has a significantly higher gastric carcinogenic rate than Western CagA (1). CagA is a multifunctional virulence factor. One of the factors contributing to the high pathogenicity of East Asian CagA is its high SHP-2-binding activity caused by a single AAC in the EPIYA-D segment (16).

Therefore, many studies have focused on the nucleotide polymorphism surrounding the EPIYA motif at the C-terminal region. However, limited information is currently available on the polymorphism at the N-terminal region related to the virulence of East Asian CagA.

In the present study, we focused on putative variants at the N-terminal region of the *cagA* sequence in *H. pylori* strains with East Asian CagA. Our WGS dataset showed four highly frequent AACs at the N-terminal region of East Asian CagA, two of which (V356A and Y677F) exhibited reproducible specificity using the validation dataset in the NCBI database.

According to previous studies, Domains I, II, and III at the N-terminal region of CagA each have their own function associated with the development of gastric carcinoma (15, 8, 35, 19). Among these 2AACs, V356A occurred in the β sheet of Domain II and Y677F in Domain III at the N-terminal region. Since Domain I is structurally isolated from Domains II and III (15), the influence of these 2 AACs on the function of Domain I may be limited. In the present study, we focused on the influence of these 2 AACs on the functions of Domains II and III. Domain II has helix α18, which binds with PS in the inner leaflet of the plasma membrane. This binding is essential for the pathophysiological effects of CagA. The binding of helix α18 with PS is dependent on electrostatic interactions and the positive charge of helix α18 is critical to the localization of CagA to the inner leaflet of the plasma membrane. The docking simulation model between CagA and PS showed that the mutant CagA reflecting both AACs did not induce a comprehensive change in ΔG, indicating that two AACs may not influence CagA binding with PS.

Domain II also has a β sheet that binds with the ectodomain of α5β1 integrin in host cells (20). Although many types of integrins are expressed in gastric epithelial cells (41), α5β1 integrin was previously reported to bind CagA with an approximately 100-fold stronger affinity than that of the natural integrin ligand, fibronectin (19). CagA attaches to bind to gastric epithelial cells and triggers its own uptake into gastric epithelial cells without infection with *H. pylori* in an endocytic manner (34). Secreted CagA has been suggested to translocate into host cells via an endocytic pathway without being injected by T4SS, and binding of CagA with PS and β1 integrin on the outer membrane of gastric cells is important in this endocytic pathway (34). Our docking simulation model between CagA and α5β1 integrin showed that the mutant CagA reflecting both AACs (V356A and Y677F) had less energy as a result of the higher affinity of CagA to α5β1 integrin, while that between CagA and PS did not show any comprehensive changes, suggesting that these two AACs in the N-terminal region of CagA have the potential to change some of the functions of CagA, such as triggering its own uptake into host cells. One of the mutations identified in our study, Y677F, is located in Domain III. Thus, it is located away from Domain II, which is the binding site of α5β1 integrin. Several previous studies demonstrated that amino acid mutation can affect the binding ability of a protein when it is located away from the interaction site. For example, van Wijk et al. reported that a point mutation slightly disturbs the free energy balance, which can subsequently alter the dynamics of the entire protein structure and change the binding affinity at a site located far from the mutation (39). Based on these observations, it is possible that the Y677F mutation caused a slight change in the structure of the protein and consequently the dynamics of Domain III, such that the binding site for α5β1 integrin became more accessible. Domain III has NBS, which bundles with CBS located at the C-terminal tail, creating a C-terminal lariat loop that strengthens the interaction between CagA and SHP2. Since the tertiary conformation of NBS-CBS binding has not yet been elucidated, a docking simulation of CagA with SHP2 by ClusPro cannot be performed. The sole AAC (Y677F) occurring in Domain III may be able to affect this function; therefore, the influence of AACs on this function needs to be evaluated in future studies. There were limitations regarding the sole use of an *in silico* model and examining only a few of the functions of the N-terminal region of CagA. Further studies involving *in vivo* and *in vitro* experiments are warranted to elucidate the relationship between these AACs and the strong pathogenesis of East Asian CagA.

In the present study, we identified four specific AACs at the N-terminal region of East Asian CagA using sequence data obtained by WGS, which we performed previously, and narrowed them down to two reproducible AACs (V356A and Y677F) using validation data. The present results provide new candidate AACs at the N-terminal region of CagA that affect the virulent function of CagA. An investigation on how these AACs influence the function of CagA is considered to be important for clarifying the mechanisms underlying the strong pathogenicity of East Asian CagA.

ACKNOWLEDGMENTS

This study was supported, in part, by a grant from JSPS KAKENHI (a Grant-in-Aid for Scientific Research) to A. I., Grant Number 15H06404, and that to J. I., Grant Number JP17K15895. The authors thank R. Okada, A. Iwamoto, and H. Ogawa for their valuable discussions. During the study, Dr. Takeshi Azuma, who was the professor of the Department of Internal Medicine, Graduate School of Medicine, Kobe University, passed away to our great regret. We sincerely appreciate his dedication to our research and hope his soul rests in peace.

REFERENCES

1. **Abe, T., Kodama, M., Murakami, K., Matsunari, O., Mizukami, K., Inoue, K., Uchida, M., Okimoto, T., Fujioka, T., Uchida, T., Moriyama, M., and Yamaoka, Y.** 2011. Impact of Helicobacter pylori CagA diversity on gastric mucosal damage: an immunohistochemical study of East-Asian-type CagA. J Gastroenterol Hepatol. **26**:688-693.
2. **Azuma, T., Yamakawa, A., Yamazaki, S., Ohtani, M., Ito, Y., Muramatsu, A., Suto, H., Yamazaki, Y., Keida, Y., Higashi, H., and Hatakeyama, M.** 2004. Distinct diversity of the *cag* pathogenicity island among *Helicobacter pylori* strains in Japan. J Clin Microbiol. **42**:2508-17.
3. **Azuma, T., Yamazaki, S., Yamakawa, A., Ohtani, M., Muramatsu, A., Suto, H., Ito, Y., Dojo, M., Yamazaki, Y., Kuriyama, M., Keida, Y., Higashi, H., and Hatakeyama, M.** 2004. Association between diversity in the Src homology 2 domain-containing tyrosine phosphatase binding site of *Helicobacter pylori* CagA protein and gastric atrophy and cancer. J Infect Dis. **189**:820-27.
4. **Azuma, T., Yamakawa, A., Yamazaki, S., Fukuta, K., Ohtani, M., Ito, Y., Dojo, M., Yamazaki, Y., and Kuriyama, M.** 2002. Correlation between variation of the 3' region of the cagA gene in Helicobacter pylori and disease outcome in Japan. J Infect Dis. **186**:1621-30.
5. **Backert, S., and Meyer, T.F.** 2006. Type IV secretion systems and their effectors in bacterial pathogenesis. Curr Opin Microbiol. **9**:207-17.
6. **Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T.** 2017. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. **7**:10480.
7. **Blaser, M.J., Perez-Perez, G.I., Kleanthous, H., Cover, T.L., Peek, R.M., Chyou, P.H., and Stemmermann, G.N.** 1995. Nomura A. Infection with Helicobacter pylori strains possessing cagA is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res. **55**:2111-15.
8. **Buti, L., Spooner, E., Van der Veen, A.G., Rappuoli, R., Covacci, A., and Ploegh, H.L.** 2011. Helicobacter pylori cytotoxin-associated gene A (CagA) subverts the apoptosis-stimulating protein of p53 (ASPP2) tumor suppressor pathway of the host. Proc Natl Acad Sci U S A. **108**:9238-43.
9. **Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R., and Covacci, A.** 1996. Cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors. Proc Natl Acad Sci U S A. **93**:14648-53.
10. **Covacci, A., Censini, S., Bugnoli, M., Petracca, R., Burroni, D., Macchia, G., Massone, A., Papini, E., Xiang, Z., and Figura, N.** 1993. Molecular characterization of the 128-kDa immunodominant antigen of Helicobacter pylori associated with cytotoxicity and duodenal ulcer. Proc Natl Acad Sci U S A. **90**:5791-95.
11. **Grosdidier, A., Zoete, V., and Michielin, O.** 2011. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. **39**:W270-77.
12. **Guex, N., and Peitsch, M.C.** 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. **18**:2714-23.
13. **Hatakeyama, M.** 2004. Oncogenic mechanisms of the Helicobacter pylori CagA protein. Nat Rev Cancer. **4**:688-94.
14. **Hatakeyama, M.** 2017. Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer. Proc Jpn Acad Ser B Phys Biol Sci. **93**:196-219.
15. **Hayashi, T., Senda, M., Morohashi, H., Higashi, H., Horio, M., Kashiba, Y., Nagase, L., Sasaya, D., Shimizu, T., Venugopalan, N., Kumeta, H., Noda, NN., Inagaki, F., Senda, T., and Hatakeyama, M.** 2012. Tertiary structure-function analysis reveals the pathogenic signaling potentiation mechanism of *Helicobacter pylori* oncogenic effector CagA. Cell Host Microbe. **12**:20-33.
16. **Hayashi, T., Senda, M., Suzuki, N., Nishikawa, H., Ben, C., Tang, C., Nagase, L., Inoue, K., Senda, T., and Hatakeyama, M.** 2017. Differential Mechanisms for SHP2 Binding and Activation Are Exploited by Geographically Distinct Helicobacter pylori CagA Oncoproteins. Cell Rep. **20**:2876-90.
17. **Higashi, H., Yokoyama, K., Fujii, Y., Ren, S., Yuasa, H., Saadat, I., Murata-Kamiya, N., Azuma, T., and Hatakeyama, M.** 2005. EPIYA motif is a membrane-targeting signal of Helicobacter pylori virulence factor CagA in mammalian cells. J Biol Chem. **280**:23130-37.
18. **Iwamoto, A., Tanahashi, T., Okada, R., Yoshida, Y., Kikuchi, K., Keida, Y., Murakami, Y., Yang, L., Yamamoto, K., Nishiumi, S., Yoshida, M., and Azuma, T.** 2014. Whole-genome sequencing of clarithromycin resistant Helicobacter pylori characterizes unidentified variants of multidrug resistant efflux pump genes. Gut Pathog. **6**:27.
19. **Jiménez-Soto, LF., Kutter, S., Sewald, X., Ertl, C., Weiss, E., Kapp, U., Rohde, M., Pirch, T., Jung, K., Retta, S.F., Terradot, L., Fischer, W., and Haas, R.** 2009. *Helicobacter pylori* Type IV Secretion Apparatus Exploits β1 Integrin in a Novel RGD-Independent Manner. PLoS Pathogens. **5**:e1000684.
20. **Kaplan-Türköz, B., Jiménez-Soto, L.F., Dian, C., Ertl, C., Remaut, H., Louche, A., Tosi, T., Haas,**

R., and Terradot, L. 2012. Structural insights into Helicobacter pylori oncoprotein CagA interaction with β1 integrin. Proc Natl Acad Sci U S A. **109**:14640-45.

21. **Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R., and Vajda, S.** 2013. How good is automated protein docking? Proteins. **81**:2159-66.
22. **Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S.** 2017. The ClusPro web server for protein-protein docking. Nat Protoc. **12**:255-78.
23. **Lind, J., Backert, S., Pfleiderer, K., Berg, D.E., Yamaoka, Y., Sticht, H., and Tegtmeyer, N.** 2014. Systematic analysis of phosphotyrosine antibodies recognizing single phosphorylated EPIYA-motifs in CagA of Western-type *Helicobacter pylori* strains. PloS ONE. **9**:e96488
24. **Murata-Kamiya, N., Kikuchi, K., Hayashi, T., Higashi, H., and Hatakeyama, M.** 2010. Helicobacter pylori exploits host membrane phosphatidylserine for delivery, localization, and pathophysiological action of the CagA oncoprotein. Cell Host Microbe. **7**:399-411.
25. **Nagae, M., Re, S., Mihara, E., Nogi, T., Sugita, Y., and Takagi, J.** 2012. Crystal structure of α5β1 integrin ectodomain: Atomic details of the fibronectin receptor. J. Cell Biol. **197**:131-40
26. **Ogawa, H., Iwamoto, A., Tanahashi, T., Okada, R., Yamamoto, K., Nishiumi, S., Yoshida, M., and Azuma, T.** 2017. Genetic variants of Helicobacter pylori type IV secretion system components CagL and CagI and their association with clinical outcomes. Gut Pathog. **9**:21.
27. **Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N., and Sibley, R.K.** 1991. Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med. **325**:1127-31.
28. **Parsonnet, J., Friedman, G.D., Orentreich, N., and Vogelman, H.** 1997. Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection. Gut. **40**:297-301.
29. **Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E.** 2011. UCSF Chimera-A visualization system for exploratory research and analysis. J Comput Chem. **25**:1605-12.
30. **Saito, Y., Murata-Kamiya, N., Hirayama, T., Ohba, Y., and Hatakeyama, M.** 2010. Conversion of Helicobacter pylori CagA from senescence inducer to oncogenic driver through polarity-dependent regulation of p21. J Exp Med. **207**:2157-74.
31. **Satomi, S., Yamakawa, A., Matsunaga, S., Masaki, R., Inagaki, T., Okuda, T., Suto, H., Ito, Y., Yamazaki, Y., Kuriyama, M., Keida, Y., Kutsumi, H., and Azuma, T.** 2006. Relationship between the diversity of the *cagA* gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan. J Gastroenterol. **41**:668-73.
32. **Suerbaum, S., and Michetti, P.** 2002. *Helicobacter pylori* infection. N Engl J Med. **347**:1175-86.
33. **The EUROGAST Study Group.** 1993. An international association between *Helicobacter pylori* infection and gastric cancer. Lancet. **341**:1359-62.
34. **Tohidpour, A., Gorrell, R.J., Roujeinikova, A., and Kwok, T.** 2017. The Middle Fragment of Helicobacter pylori CagA Induces Actin Rearrangement and Triggers Its Own Uptake into Gastric Epithelial Cells. Toxins. **28**;9:(8).
35. **Tsang, Y.H., Lamb, A., Romero-Gallo, J., Huang, B., Ito, K., Peek, R.M., Ito, Y., and Chen, L.F.** 2010. Helicobacter pylori CagA targets gastric tumor suppressor RUNX3 for proteasome-mediated degradation. Oncogene. **29**:5643-50.
36. **Ulloa-Guerrero, C.P., Delgado, M.D.P., and Jaramillo, C.A.** 2018 Structural Analysis of Variability and Interaction of the N-terminal of the Oncogenic Effector CagA of *Helicobacter pylori* with Phosphatidylserine. Int J Mol Sci. **19**:3273.
37. **Vajda, S., Yueh, C., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R., and Kozakov, D.** 2017. New additions to the ClusPro server motivated by CAPRI. Proteins. **85**:435-44.
38. **Wotherspoon, A.C., Ortiz-Hidalgo, C., Falzon, M.R., and Isaacson, P.G.** 1991. Helicobacter pylori-associated gastritis and primary B-cell gastric lymphoma. Lancet. **338**:1175-76.
39. **van Wijk, S.J., Melquiond, A.S., de Vries, S.J., Timmers, H.T., and Bonvin, A.M.** 2012. Dynamic Control of Selectivity in the Ubiquitination Pathway Revealed by an ASP to GLU Substitution in an Intra-Molecular Salt-Bridge Network. PLoS Comput Biol. **8**;11:e1002754.
40. **Yamazaki, S., Yamakawa, A., Okuda, T., Ohtani, M., Suto, H., Ito, Y., Yamazaki, Y., Keida, Y., Higashi, H., Hatakeyama, M., and Azuma, T.** 2005. Distinct diversity of vacA, cagA, and cagE genes of Helicobacter pylori associated with peptic ulcer in Japan. J Clin Microbiol. **43**:3906-16.
41. **Zhao, Q., Busch, B., Jiménez-Soto, L.F., Ishikawa-Ankerhold, H., Massberg, S., Terradot, L., Fischer, W., and Haas, R.** 2018. Integrin but not CEACAM receptors are dispensable for Helicobacter pylori CagA translocation. PLoS Pathog. **26**;14:10.