Variations in cag pathogenicity island genes of Helicobacter pylori from Latin American groups may influence neoplastic progression to gastric cancer

Cosmeri Rizzato1*, Javier Torres2, Ofure Obazee3, Margarita Camorlinga-Ponce2, Esperanza Trujillo4, Angelika Stein3, Alfonso Mendez-Tenorio5, Maria Mercedes Bravo4, Federico Canzian5,6 & Ikuko Kato6

Helicobacter pylori (Hp) colonizes the human stomach and induces acute gastritis, peptic ulcer disease, atrophic gastritis, and gastric adenocarcinoma. Increased virulence in Hp isolates derives from harboring the cag (cytotoxin-associated genes) pathogenicity island (cagPAI). We analyzed the microvariants in cagPAI genes with the hypothesis that they may play an important role in determining Hp virulence. We tested DNAs from cagA positive patients Hp isolates; a total of 74 patients with chronic gastritis (cG, n = 37), intestinal metaplasia (IM, N = 21) or gastric cancer (GC, N = 16) from Mexico and Colombia. We selected 520 non-synonymous variants with at least 7.5% frequency in the original sequence outputs or with a minimum of 5 isolates with minor allele. After adjustment for multiple comparisons, no variants were statistically significantly associated with IM or GC. However, 19 non-synonymous showed conventional P-values < 0.05 comparing the frequency of the alleles between the isolates from subjects with gastritis and isolates from subjects with IM or GC; 12 of these showed a significant correlation with the severity of the disease. The present study revealed that several cagPAI genes from Latin American Western Hp strains contains a number of non-synonymous variants in relatively high frequencies which could influence on the clinical outcome. However, none of the associations remained statistically significant after adjustment for multiple comparison.

Gastric cancer has the third highest mortality rate and the fifth highest incidence worldwide1. The two regions of the world with highest mortality rate for gastric cancer are Asia and Latin America accounting for almost two thirds of all gastric cancer deaths2. Within the US, ethnic minorities, e.g., Asians, Blacks, Hispanics and Native Americans, experience an incidence almost twice as high as non-Hispanic Whites3. Some Asian countries, such as Japan, have nation-wide screening programs for early detection of gastric cancer, whereas most other high-risk countries, such as Latin American countries, do not have such programs, nor does the US for ethnic minorities4–6. Although Helicobacter pylori (HP) is an established cause of gastric cancer7, eradication of HP in general asymptomatic populations has not been advised, because of the large number of persons already infected in high risk populations (>90%), high reinfection rates in endemic areas8, antibiotic resistance, high cost of the treatment9 and the increased risk of esophageal cancer associated with HP-negative/eradicated individuals10–12. Thus, new strategies for gastric cancer prevention are warranted and may help reduce health disparities, mainly in the most affected and underdeveloped regions of the world13. All considered identification of new HP variants potentially

1Department of Translation Research and of New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy. 2Unidad de Investigación en Enfermedades Infecciosas, UMAE Pediatría, Instituto Mexicano del Seguro Social, Mexico City, Mexico. 3Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. 4Grupo de Investigación en Biología del Cáncer. Instituto Nacional de Cancerología, Bogotá, Colombia. 5Laboratorio de Biotecnología y Bioinformática Genómica, ENCB, Instituto Politécnico Nacional, México City, México. 6Department of Oncology and Pathology, Wayne State University School of Medicine, Detroit, MI, USA.

*email: cosmeri.rizzato@unipi.it
HP has a remarkably high level of genetic diversity due to recombination rates higher than in any other known bacterial species7,14,15. A number of HP virulence factors have been identified, but it is now clearly established that cagA and the cytotoxin-associated gene pathogenicity island (cagPAI) play a central role in the pathogenesis of HP-associated diseases16,17. The cagPAI consists of a 40 kb chromosomal DNA and is present in approximately 95% of East Asian isolates, compared to 60% of low-risk Western isolates18–21. These genes encode cytotoxins and components of the type IV secretion system (T4SS) that acts as a molecular syringe injecting bacterial macromolecules into host cell cytosol18. This ultimately leads to sustained IL-8 production, inflammation, proliferation and morphological changes of gastric epithelial cells which underlie HP-induced gastric premalignant and malignant pathologies19. However, the presence of cagA (a marker of cagPAI) does not predict clinical outcomes in high-risk populations since the majority of HP are cagA positive strains, and among the infected subject less than 3% develop gastric cancer.

CagPAI contains 31 open reading frames, named cag1 to cag26 or cagA to cagZ and cagO to cagO, or by locus name of the HP 26695 or HP strains genomes22. A number of the cagPAI genes are homologous to type IV secretion system genes (T4SS) of Agrobacterium, VirB1-11, VirD4, and VirE1. In Agrobacterium tumefaciens model, the T4SS is composed of 13 proteins, which span both bacterial and host cell membranes23–26. Because of the relevance of cagPAI in the biological activities of HP that lead to tissue damage, microvariations in cagPAI genes other than cagA are likely to play an important role in determining HP virulence. In this work we aim to study the microvariability of all the cagPAI genes with the exception of cagY, for which we limited our analysis to 339 nucleotides at the 3’ end of the genes encoding for the last carboxy terminal 113 amino acids, CagYe27. The present study is meant to extend our previous work based on 454 sequencing28 by including new cases with premalignant lesions, and by studying additional cagPAI genes using a newer sequencing platform to increase the power of SNP detection. Data from the present work have been included in a phylogenetic analysis of Latin American HP strains29. We performed a comprehensive screening of single nucleotide polymorphisms (SNP) in relation to gastric histopathology in order to identify variants with potential predictive value for clinical outcome, which warrant further validation in larger samples as well as separate functional analyses.

Results

Performance of genome-wide sequencing. We prepared sequence libraries and performed whole-genome sequencing on 92 HP strains, including reference strain 26695. Details of the sequencing outcome are included in supplementary table 1. We obtained a total number of reads per samples between 53,427 and 1,366,065. Genomes were assembled, and we found that the percentage of reads that aligned to the reference sequence ranged between 0% and 60.51%. We excluded from further analysis 3 genomes for which no reads could be aligned, 9 where cagPAI was absent, 2 strains isolated from the same patient and reference strain 26695 that was analyzed only as quality control (data not shown).

The coverage of the whole genome for the 74 samples selected for further analysis was on average 39.52× (range 6.31×–114.45×), and coverage for the cagPAI was 85.31× on average (range 11.08×–466.87×). The number of reads that aligned to the cagPAI per sample ranged between 3201 and 43462. cagA gene was missing in 11 of 92 sequenced strains and cag1 in one strain.

Variability by gene. We summarized the genetic variability detected in the twenty-six cagPAI genes in Table 1. For each gene we computed the degree of variability as the number of sites showing a different genotype compared with the reference strain out of the total number of nucleotides in the gene, both as synonymous and non-synonymous variations (causing therefore differences at amino acid level). We assessed a range of variability from 9.54% in cagF to 31.22% in cagA at DNA level calculated as ratio of polymorphic position to gene length, while the amino acid variability ranged from 1.8% in the cagE gene, which is the minimum variation that we found (with the exception of the analyzed region of cagY in which we did not find non-synonymous variations), to 17.82% in cagC calculated as ratio of non-synonymous polymorphisms to number of amino acid in the translated protein. The number of polymorphisms identified for each gene are summarized in Table 1.

Comparison of frequencies of polymorphisms showing a differential distribution between gastritis and IM or GC cases. When comparing the frequency of the 520 selected alleles between the isolates from subjects with gastritis and isolates from subjects with IM or GC we found statistically significant differences in allele distribution for three polymorphisms in cagA gene (Q/K427R, N467G and V1041T), three in cagC gene (V22I, V37I, 145V), one in the cagE gene (K981E), one in cagL gene (S10F), one in cagX gene (G11N), one in cagS (G146D), one in cagY (S35A), three in cagO (V353I, P406L, N407E) and one in cagZ (N125A) (Table 2). Furthermore 4 polymorphisms in cagA (V521, G656R, S194F and Q/R427K) showed a significant trend with grade of the disease.

When IM and GC were analyzed separately and compared with the non-atrophic gastritis, 7 SNPs showed a marginally (P < 0.05) statistically significant association when comparing gastritis vs. IM and 10 when comparing gastritis vs. GC (Table 3). In the cagA gene 3 of these variants were associated with risk of IM (V521, S194F and Q/R427K) and one with GC (N467G), as shown in Table 3. In the cagC gene we detected three SNPs with a differential distribution between gastritis and GC (V22I, V37I, 145V, see Table 3). In cagL gene polymorphism (S10F) showed a differential allelic distribution between isolates derived from IM cases and gastritis cases (OR = 0.14; 95% CI 0.04–0.46, P = 0.002) (Fig. 1, Table 3). For the cagX gene polymorphism G11N showed a differential distribution between cases of IM and gastritis (OR = 0.20; 95% CI 0.06–0.71, P = 0.011) (Fig. 1, Table 3). For cagS one polymorphism (S35A) showed a differential allelic distribution between isolates derived from IM cases and gastritis cases (OR = 8.80; 95% CI 2.29–33.84, P = 0.001) (Fig. 1, Table 3). In cagO two adjacent polymorphisms
showed a differential distribution between cases of cancer cases and gastritis in particular the association for P406L showed an OR \(= \) \(7.97, 95\% \text{ CI } 2.03–31.27\) and for N407E OR \(= \) \(10.29, 95\% \text{ CI } 2.45–43.15\). Additionally, the allelic distributions of these two polymorphisms were significantly different between Mexican and Colombian samples (data not shown), namely the variant allele frequencies were extremely low in Colombia, therefore the associations were driven by the Mexican cancer cases.

In the \(cag\)β gene variant N125A showed an inverse association with cancer with an OR of 0.14 (95\% CI 0.03–0.66, \(P = 0.013\)) (Fig. 1, Table 3).

Next, a multiple comparison analysis was performed by applying a Bonferroni-corrected threshold, and none of the above described SNPs showed a \(P\)-value lower than the threshold adjusted for this type of analysis of \(P = 9.6 \times 10^{-5} \) \((0.05/520)\). None of the SNPs reached this study-wise \(P\)-value. Supplementary table 2 lists all the polymorphisms observed in 24 analyzed genes.

**Table 1.** Overview of genetic variability in genes in HP \(cag\)PAI. \(^a\)Number of polymorphic positions differ from number of variants because we found that several indel variants span more than one nucleotide. \(^b\)Non-synonymous variants with at least 7.5\% frequencies when compared to the reference sequence.

We also analyzed EPIYA (A, B or C) and CM (cm) motifs distribution in 72 \(cag\)A positive sequenced strains\(^{30–32}\), while two strains were \(cag\)PAI positive but lacking the \(cag\)A gene. We found a high degree of variability consisting of 12 different patterns, all of the Western Type (supplementary figure 1). Most of the strains (50) presented the A/B/cm/C/cm pattern, followed by the pattern A/B/cm/A/bcm/C/cm (in 9 strains), and the pattern A/bcm/A/bcm/C/cm (in 3 strains). Other less frequent patterns are described in supplementary figure 1. There was no difference in the distribution of the various patterns when compared between the three different disease groups.

**Discussion**

This study was conducted in Latin American HP strains, in order to identify specific \(cag\)PAI micro-variants associated with high-grade gastric lesions. This effort is of high clinical and translational importance as the presence of \(cag\)A gene does not predict outcomes of HP infection, particularly in high-risk populations where the majority of the strains carry \(cag\)PAI. The present study not only confirmed the extremely high variability in the \(cag\)PAI genes, but also pointed to several variants with potential clinical relevance in a few genes for future studies.

Other study have investigated the whole \(cag\)PAI\(^{19}\) however none have performed an extensive analysis of polymorphic variant of each gene in the region.
cagA variants in find that carriers of D at this position are at lower risk of GC in comparison with the asparagine (N) carriers. In our previous publication we did not observe the Y amino acid in position 58 in any sample, although we did more efficiently a shift of gastric integrin a5b1 in the corpus, which has been related with gastric carcinogenesis. Acid (E) in position 59 (Y58E59) compared with the combination aspartic acid (D58) and Lysine (K59), induced sary for the formation of HP pilus40, mediating the stabilization of cagT which is a T4SS structural protein 41,42.

glycine (non-polar) to asparagine (polar). cagX, Vir9 homologous protein, has been found recently to be neces-

The other variant that is considered to be detrimental non-tolerant is located at residue 11 of cagX suggesting that this is a Western-strain specific variant35. Thus, overall, the differences observed in these studies are likely to be driven by geographical origins of HP. The other variant that is considered to be detrimental non-tolerant is located at residue 11 of cagX, exchanging glycine (non-polar) to asparagine (polar), cagX, Vir9 homologous protein, has been found recently to be necessary for the formation of HP pilus40, mediating the stabilization of cagT which is a T4SS structural protein41,42.

| Gene   | Nucleotide change | Amino acid change | IM + GCa | Gastritis casesa | OR (95%CI) | FisherP-value |
|--------|-------------------|------------------|----------|-----------------|------------|--------------|
| cagA   | G154A             | V52I             | 0.17     | 0               | 14.45 (0.76–273.50) | 0.02        |
| cagA   | G193A             | G65R             | 0.2      | 0.03            | 8.00 (0.90–76.92)   | 0.047       |
| cagA   | C581T             | S194F            | 0.17     | 0               | 14.45 (0.76–273.50) | 0.02        |
| cagA   | A1280G            | Q/K427R          | 0.97     | 0.7             | 12.61 (1.50–105.81) | 0.007       |
| cagA   | C1279A            | Q/R427K          | 0        | 0.21            | 0.06 (0.00–1.06)    | 0.011       |
| cagA   | AA1399–1400GG     | N467G            | 0.73     | 0.42            | 3.73 (1.29–10.81)   | 0.021       |
| cagA   | GTT/CCC3121–3123ACC | Y/P1041T | 0.57     | 0.3             | 3.01 (1.07–8.47)    | 0.044       |
| cagC   | G64A              | Y22I             | 0.38     | 0.14            | 3.90 (1.23–12.34)   | 0.032       |
| cagC   | G109A             | Y37I             | 0.35     | 0.11            | 4.47 (1.30–15.41)   | 0.025       |
| cagC   | A133G             | I45V             | 0.3      | 0.08            | 4.79 (1.21–18.96)   | 0.035       |
| cagE   | A2941G            | K981E            | 0.08     | 0.32            | 0.18 (0.05–0.72)    | 0.019       |
| cagL   | C29T              | S10F             | 0.51     | 0.78            | 0.29 (0.11–0.80)    | 0.027       |
| cagX   | GG31–32AA         | G11N             | 0.24     | 0.54            | 0.27 (0.10–0.74)    | 0.017       |
| HP0520_cagL | T103G            | S35A             | 0.32     | 0.11            | 3.84 (1.1–13.36)    | 0.046       |
| HP0522_cagA | G1057A           | Y35S1            | 0.78     | 0.56            | 2.9 (1.04–8.06)     | 0.048       |
| HP0522_cagE | C1217T           | P406L            | 0.38     | 0.14            | 3.77 (1.19–11.98)   | 0.032       |
| HP0522_cagA | AAT1219–1222GAG    | N407E            | 0.38     | 0.11            | 4.87 (1.42–16.72)   | 0.013       |
| HP0524_cag3 | AAT373–375GCA     | N125A            | 0.69     | 0.92            | 0.21 (0.05–0.82)    | 0.035       |
| HP0534_cagS | GC437–438AT       | G146D            | 0.84     | 0.60            | 3.44 (1.14–10.4)    | 0.035       |

Table 2. Polymorphisms in cagPAI genes showing a differential distribution between gastritis and IM + GC cases. aFrequency of variant alleles.

Overall sequence variability derived from this study for the selected cagPAI genes was consistent with that reported previously using different sequencing techniques.28,34, and extend the study to a higher number of genes. These results support the signature of diversifying selection through bacterial evolution in the proteins that are surface-exposed and involved in interactions with host molecules. However, the frequencies of amino acid variants in cagA, cagC and cagE found in this study were substantially higher than those previously reported.28,34. This may be partially due to a greater number of strains under investigation in our work compared to previous publications. While we cannot completely rule out artifacts from this high throughput platform, such artifacts should affect both synonymous and non-synonymous variants.

In a previous study conducted with amplicon sequencing by 454 for 84 Mexican and 11 Venezuelan samples we reported 10 non-synonymous SNPs with differential allelic distribution between gastritis and gastric cancer at conventional P-values between 0.01–0.05. In the present project that included equal numbers of Mexican and Colombian strains we did not see any disease association with these 10 variants. Although variant frequencies were not markedly different between Mexico and Colombia strains, particularly for variants showing a significant association with IM + GC, we have previously reported important phylogenetic differences between strains of these two countries. These phylogenetic differences may partly explain discordant results between the previous and current studies.

Previous publications reported an association of GC with the presence of variants in position 58 and 59 of cagL, protein; in two studies35,36 the concurrent presence of tyrosine (Y) in amino acid position 58 and glutamic acid (E) in position 59 (Y58E59) compared with the combination aspartic acid (D58) and Lysine (K59), induced more efficiently a shift of gastric integrin a5b1 in the corpus, which has been related with gastric carcinogenesis. In our previous publication we did not observe the Y amino acid in position 58 in any sample, although we did find that carriers of D at this position are at lower risk of GC in comparison with the asparagine (N) carriers.28. In the current work we confirmed the absence of polymorphism in Y position 58 and the presence of N58D polymorphism. We also observed the E59K polymorphism but we did not find association with either IM or GC in our populations. It should be noted that there was a major difference between our current and previous studies in the composition of geographical origins of the samples. Our former study included Colombian while the current study included Venezuelan strains; this is relevant considering our recent report where we document adaption of HP genome to different Latin American populations. Furthermore, Gorrel and co-worker performed an expanded analysis of this region analyzing the sequence from amino acid 58 to 62 and found significant differences according to the geographical origin; in this sense, we confirmed the predominance of the DKMGF aminoacid sequence.

Some of the variants may warrant further studies as the SIFT program predicts them to be damaging non-tolerant. In particular the cagL S10F variant (located at codon 10) changes from serine (polar) to phenylala-nine (non-polar). Interestingly, among the 43 Asian strains recently sequenced, no single S10F variant was found, suggesting that this is a Western-strain specific variant. Thus, overall, the differences observed in these studies are likely to be driven by geographical origins of HP.
Thus, cagX mutants prevent cagA biological activities\(^40\). In this context, our finding of a protective effect of the N-terminus of cagA is compatible with a lesser virulent behavior.

One cagC variant, located at codon 22, replacing valine with isoleucine, was associated with risk of gastric cancer and predicted to be intolerant by SIFT. Valine to isoleucine substitutions have been reported to result in changes in protein structure, kinetics and stability in both bacteria\(^{43-45}\) and humans\(^{45-47}\). However, the possible function of this polymorphism remains unclear.

The four polymorphisms in genes cag\(c_\delta\), cag\(c_\zeta\), cag\(c_\delta\), and cag\(c_\zeta\) showing a differential distribution in IM or cancer cases were predicted to be tolerated by SIFT.

There were several variants at the N-terminal of cagA that showed rather strong (OR > 4.60) associations with high grade lesions (\(P = 0.0025\) (Table 2) as well as a significant trend by type of lesions (\(P < 0.05\) (Table 3). Some of these significant changes in amino acid characteristics (e.g., Q427R, N467G), although none were predicted to be detrimental by SIFT. Valine to isoleucine substitutions have been reported to result in changes in protein structure, kinetics and stability in both bacteria\(^{43-45}\) and humans\(^{45-47}\). However, the possible function of this polymorphism remains unclear.

The four polymorphisms in genes cag\(c_\delta\), cag\(c_\zeta\), cag\(c_\delta\), and cag\(c_\zeta\) showing a differential distribution in IM or cancer cases were predicted to be tolerated by SIFT.

| Gene       | Nucleotide change | Amino acid change | Frequency in controls | Frequency in IM | OR (CI) | Frequency in cancer | OR (CI) | P-value for trend |
|------------|-------------------|-------------------|-----------------------|-------------------|----------|---------------------|----------|------------------|
| cagA       | G154A             | V52I              | 0                     | 0.17              |          | 15.13 (0.84–311.2)\(^*\) | 0.17     | 16.0 (0.71–359.3) | 0.029 |
| cagA       | G193A             | G65R              | 0.03                  | 0.17              |          | 6.40 (0.61–66.76) | 0.25     | 10.67 (0.99–115.36) | 0.026 |
| cagA       | C851T             | S194F             | 0                     | 0.22              |          | 20.79 (1.05–411.83) | 0.08     | 8.74 (0.33–229.93) | 0.110 |
| cagA       | A1280G            | Q/K427R           | 0.7                   | 1                 |          | 16.53 (0.91–300.97)\(^*\) | 0.92     | 4.78 (0.54–42.21) | 0.024 |
| cagA       | CI279A            | Q/R427K           | 0.21                  | 0                 |          | 0.10 (0.01–1.78)\(^*\) | 0        | 0.14 (0.01–2.68) | 0.017 |
| cagA       | AI2399–1400GG     | N467G             | 0.42                  | 0.67              | 2.71 (0.82–9.00) | 0.83     | 6.79 (1.28–35.97) | 0.010 |
| cagA       | GTT/CC33121–3123ACC | VP/1014T          | 0.3                   | 0.67              |          | 4.60 (1.35–15.73)\(^*\) | 0.42     | 1.64 (0.42–6.44) | 0.190 |
| cagC       | G64A              | V22I              | 0.14                  | 0.19              | 1.51 (0.35–6.36) | 0.63     | 10.67 (2.68–42.53) | 0.001 |
| cagC       | G190A             | V37I              | 0.11                  | 0.33              |          | 4.13 (1.04–16.37) | 0.38     | 4.95 (1.16–21.09) | 0.019 |
| cagC       | AI33G             | I45V              | 0.08                  | 0.19              |          | 2.67 (0.54–13.29) | 0.44     | 8.81 (1.89–41.08) | 0.003 |
| cagE       | A2941G            | K981E             | 0.32                  | 0.05              |          | 0.10 (0.01–0.87)\(^*\) | 0.13     | 0.30 (0.06–1.52) | 0.039 |
| cagL       | C29T              | S10F              | 0.78                  | 0.33              | 0.14 (0.04–0.46)\(^*\) | 0.75     | 0.83 (0.21–3.28) | 0.310 |
| cagX       | GG31–32AA         | G11N              | 0.54                  | 0.19              | 0.20 (0.06–0.71)\(^*\) | 0.31     | 0.39 (0.11–1.33) | 0.040 |
| HP0520_cagA| T103G             | S35A              | 0.11                  | 0.52              | 8.80 (2.29–33.84)\(^*\) | 0.0625   | 0.53 (0.05–5.19) | 0.626 |
| HP0522_cagA| G1057A            | V353I             | 0.56                  | 0.81              | 3.40 (0.95–12.13) | 0.75     | 2.40 (0.65–8.88) | 0.093 |
| HP0522_cagA| CI217T            | P406L             | 0.14                  | 0.24              | 1.94 (0.49–7.69) | 0.5625   | 7.97 (2.03–31.27) | 0.0023 |
| HP0522_cagA| AAT1219–1222GAG   | N407E             | 0.11                  | 0.24              | 2.50 (0.59–10.61) | 0.5625   | 10.29 (2.45–43.15) | 0.0008 |
| HP0524_cagA| AAT373–375GCA     | N125A             | 0.92                  | 0.76              | 0.29 (0.06–1.37) | 0.6       | 0.14 (0.03–0.66) | 0.008 |
| HP0534_cagA| GG437–438AT       | G146D             | 0.60                  | 0.81              | 2.83 (0.79–10.21) | 0.875    | 4.67 (0.92–23.79) | 0.028 |

Table 3. Polymorphisms in cagPAI genes showing a differential distribution between gastritis and IM or GC cases. *\(P < 0.05\) by Fisher Exact test.
Methods

Study population. Strains analyzed in this study were isolated from patients recruited in the context of a multi-centric study based in Latin America\(^{50,51}\). Sequences used in the present work are largely overlapping with those reported in a phylogenetic analysis on Latin American HP strains\(^{29}\). Patients attended the gastroenterology or oncology services and were subjected to endoscopy for diagnostic purposes. We isolated HP in 92 of these patients and sequenced them, however 18 were dropped due to poor quality or because they did not carry the \textit{cag}PAI. Samples included in the following analysis were therefore 74 HP clinical isolates from 74 individuals recruited in Colombia (N = 37) and Mexico (N = 37, nine of which have been already sequenced with the 454 technology for 5 genes\(^{28}\)). Thirty-seven of these subjects had non-atrophic or atrophic gastritis, 21 intestinal metaplasia (IM) and 16 distal gastric cancer (GC). Table 4 shows pertinent characteristics of the population. For Mexican samples, all the patients signed an informed consent and the study was approved by ethical committees of the Instituto Mexicano del Seguro Social (IMSS) and General Hospital of the Secretaria de Salud (SS), Mexico City, Mexico\(^{28}\). For Colombia, the clinical studies where patients were originally recruited were approved by the Ethical and Research Committee of the Instituto Nacional de Cancerología, and all the patients signed an informed consent. This study was approved by the Ethical and Research Committee of the Instituto Nacional de Cancerología\(^{52}\). All research was performed in accordance with relevant guidelines and regulations.

Sample preparation. In order to isolate HP, Mexican stomach biopsies were homogenated and inoculated on 5% sheep Blood agar base (Becton Dickinson, New Jersey, USA) supplemented with vancomycin, trimethoprim, polymyxin B (Campylobacter-selective antibiotics, Oxoid, LTD. England). Colombian biopsies were homogenated and cultured on blood agar plates, supplemented with Campylobacter-selective supplement (Oxoid), 1% Vitox (Oxoid), 7% horse serum (Invitrogen). Plates, in both centers, were incubated at 37 °C under a 10% CO\(_2\) atmosphere, and genomic DNA was extracted from HP colonies using the DNeasy Mini Kit (Qiagen, Hilden Germany).

Whole-genome HP sequencing. Nextera XT sample preparation kit (Illumina) was used for preparation of the sequencing libraries according the manufacturer’s instructions. 1 ng of dsDNA libraries was quantified with Picogreen and used as input for Illumina sequencing. High Sensitivity DNA Kit (Agilent Technologies) was used to verify the fragment length distribution of the libraries using the agilent Bioanalyzer. Sequencing was performed on an Illumina HiSeq. 2500 sequencer with v4 PE125 chemistry at the Genomics and Proteomics Core Facility of DKFZ.

Bioinformatic and statistical methods. Raw sequences with Phred score >30 were analyzed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to further assess quality. The resulting sequencing output, in fastq format, was assembled with the “map by reference” function of the Geneious software platform (http://www.geneious.com/), considering the sequence of the 26695 strain (NC_000915.1) as reference. A consensus sequence was determined for \textit{cagA} (HP0547), \textit{cagC} (HP0546), \textit{cagE} (HP0544), \textit{cagF} (HP0543), \textit{cagI} (HP0540), \textit{cagL} (HP0539), \textit{cagT} (HP0527) (339 nucleotides at the 3’ end of the genes encoding for the last carboxyl terminal 113 amino acids of \textit{CagY}, \textit{cagYc})\(^{28}\), \textit{cagX} (HP0528), \textit{cagγ} (HP0523), and single nucleotide polymorphisms and small insertions and deletions were identified for each gene. For the
| Diagnosis          | Number of samples | Colombian | Mexican | Total |
|--------------------|-------------------|-----------|---------|-------|
| Gastritis cases    | 21                | 16        | 37      |       |
| Metaplasia cases   | 12                | 9         | 21      |       |
| Cancer cases       | 4                 | 12        | 16      |       |
| Gender             |                   |           |         |       |
| Female             | 13                | 24        | 37      |       |
| Male               | 22                | 13        | 35      |       |
| Unknown            | 2                 | 2         |         |       |
| Median age (25%–75%) |                 |           |         |       |
| Gastritis cases    | 43 (37–59)        | 45.5 (37.5–55.5) |       |
| Metaplasia cases   | 51.5 (46–59)      | 58 (52–68) |       |
| Cancer cases       | 70 (62.5–70)      | 53 (46–59) |       |

Table 4. Characteristics of the study population.

The analysis of cagA gene an additional strategy was used to better analyze the EPIYA and CM motifs (C-MET motif mediate CagA multimerization and membrane targeting)\(^\text{30,31}\). Illumina reads of the cagA gene were extracted and realigned by the "de novo assembly" option of the same software. To exclude potential artifacts in sequencing and to enrich variants with clinical relevance, we selected a total of 520 non-synonymous variants with at least 7.5% frequency in the HP isolates we included in the analysis or with a minimum of 5 isolates with the variant allele. A Bonferroni-corrected threshold \((P = 0.05/520 = 9.6 \times 10^{-5})\) was used to adjust for multiple comparisons. The variant alleles were determined using strain 26695 as reference. When more than one variant resulted in substitution of the same amino acid, we also analyzed the frequencies of the combined amino acid variant. We used non-atrophic gastritis as the control group to compare frequencies of variants in IM and GC groups or a combined group with the two pathologies, and genotypes at a given locus were dichotomized, a selected variant vs. all other genotypes. \(P\)-values for differences in allelic frequencies between the control and IM/GC combined or separately were determined by the Fisher’s exact test (2-sided). We also computed odds ratios (OR) 95% confidence interval (CI) for these gastric pathologies using logit estimators to obtain the CI even for the variants with zero frequencies in any category. For the variants that showed an unadjusted Fisher \(P\)-value of < 0.05, we further tested linear trends of their frequencies across the three histological groups, control, IM and GC, using Mantel-Haenszel chi-square test. These analyses were performed by SAS version 9. The effect of DNA polymorphisms on the predicted proteins was evaluated with a bioinformatics tool: SIFT (Sorting Intolerant From Tolerant) http://sift.jcvi.org\(^\text{53}\).

Data availability

All sequences from Mexican strains are deposited at DDBJ/ENA/GenBank under the bioprojects PRJNA338771 and PRJNA203445. Genome sequences from Colombia are deposited under the bioproject PRJNA352848. GenBank accession numbers for 69 out of 74 analyzed strains are reported in supplementary table 1, submission of the remaining strains to GenBank is ongoing. In the meantime, the data are available upon request to the corresponding author.

Received: 31 July 2018; Accepted: 20 March 2020;
Published online: 16 April 2020

References

1. Torre, L. A. et al. Global cancer statistics, 2012. CA Cancer J. Clin 65, 87–108, https://doi.org/10.3322/caac.21262 (2015).
2. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359–386, https://doi.org/10.1002/ijc.29210 (2015).
3. Howlader, N. et al. Vol. 2017 (eds N. Howlader et al.) (National Cancer Institute, Bethesda, MD, 2017).
4. Hamashima, C. Current issues and future perspectives of gastric cancer screening. World journal of gastroenterology 20, 13767–13774, https://doi.org/10.3748/wjg.v20.i38.13767 (2014).
5. Leung, W. K. et al. Screening for gastric cancer in Asia: current evidence and practice. Lancet Oncol 9, 279–287, https://doi.org/10.1016/S1470-2045(08)70072-X (2008).
6. Pasechnikov, V., Chukov, S., Fedorov, E., Kikuste, I. & Leja, M. Gastric cancer: prevention, screening and early diagnosis. World journal of gastroenterology 20, 13842–13862, https://doi.org/10.3748/wjg.v20.i38.13842 (2014).
7. IARC. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Vol. 83 (International Agency for Research on Cancer, 2004).
8. Morgan, D. R. et al. Risk of recurrent Helicobacter pylori infection 1 year after initial eradication therapy in 7 Latin American communities. JAMA 309, 578–586, https://doi.org/10.1001/jama.2013.311 (2013).
9. Moayyedi, P. et al. Systematic review and economic evaluation of Helicobacter pylori eradication treatment for non-ulcer dyspepsia. Dyspepsia Review Group. BMJ 321, 659–664, https://doi.org/10.1136/bmj.321.7262.659 (2000).
10. Islam, F. & Kamangar, F. Helicobacter pylori and esophageal cancer risk: a meta-analysis. Cancer Prev Res (Phila) 1, 329–338, https://doi.org/10.1158/1940-6207.CAPR-08-0109 (2008).
11. Lee, Y. C. et al. The benefit of mass eradication of Helicobacter pylori infection: a community-based study of gastric cancer prevention in South Korea. Gut 62, 676–682, https://doi.org/10.1136/gutjnl-2012-302240 (2013).
12. Ma, J. L. et al. Fifteen-year effects of Helicobacter pylori, garlic, and vitamin treatments on gastric cancer incidence and mortality. Journal of the National Cancer Institute 104, 488–492, https://doi.org/10.1093/jnci/djs003 (2012).
13. Ennis, S., Rios-Vargas, M. & Albert, N. The Hispanic Population: 2010. 2010 Census Briefs (2011).
14. Fischer, W. et al. Strain-specific genes of Helicobacter pylori: genome evolution driven by a novel type IV secretion system and genomic island transfer. Nucleic acids research 38, 6089–6101, https://doi.org/10.1093/nar/gkq378 (2010).
15. Portal-Celhay, C. & Perez-Perez, G. I. Immune responses to Helicobacter pylori colonization: mechanisms and clinical outcomes. Clin Sci (Lond) 110, 305–314, https://doi.org/10.1042/CS20050232 (2006).
51. Silva-Zolezzi, I. et al. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America* 106, 8611–8616, https://doi.org/10.1073/pnas.0903045106 (2009).

52. Gutierrez-Escobar, A. J., Trujillo, E., Acevedo, O. & Bravo, M. M. Phylogenomics of Colombian Helicobacter pylori isolates. *Gut Pathog* 9, 52, https://doi.org/10.1186/s13099-017-0201-1 (2017).

53. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 4, 1073–1081, https://doi.org/10.1038/nprot.2009.86 (2009).

**Acknowledgements**

We thank the unit for High Throughput Sequencing of DKFZ Genomics and Proteomics Core Facility led by Dr. Stephan Wolf for their valuable services in providing the sequencing data used in this project. This work was supported by the National Institutes of Health, United States [5R21CA182822 to IK].

**Author contributions**

Conceived and designed the experiments: C.R. I.K. F.C. J.T. Performed the experiments: C.R. O.O. A.S. Analyzed the data: C.R. I.K. A.M.T. Contributed reagents/ materials/analysis tools: J.T. M.C.-P. E.T. M.M.B. Wrote the paper: C.R. I.K. F.C. J.T. Revised the manuscript: A.M.T. M.M.B.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-63463-0.

**Correspondence** and requests for materials should be addressed to C.R.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020