Received: 23 October 2019 | Revised: 1 April 2020 | Accepted: 9 April 2020

DOI: 10.1002/ijc.33032

IJC International Journal of Cancer

CANCER EPIDEMIOLOGY

Genetic polymorphisms in the cag pathogenicity island of *Helicobacter pylori* and risk of stomach cancer and high-grade premalignant gastric lesions

Federico Canzian[1] | Cosmeri Rizzato[2] | Ofure Obazee[1] | Angelika Stein[1] | Lourdes Flores-Luna[3] | Margarita Camorlinga-Ponce[4] | Alfonso Mendez-Tenorio[5] | Jorge Vivas[6] | Esperanza Trujillo[7] | Hyejong Jang[8] | Wei Chen[8] | Elena Kasamatsu[9] | Maria Mercedes Bravo[7] | Javier Torres[4] | Nubia Muñoz[10] | Ikuko Kato[8]

[1]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Department of Translation Research and of New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy

[3]Center for Public Health Research, National Institute of Public Health, Cuernavaca, Morelos, Mexico

[4]Unidad de Investigación en Enfermedades Infecciosas, UMAE Pediatría, Instituto Mexicano del Seguro Social, Mexico City, Mexico

[5]Laboratorio de Biotecnología y Bioinformática Genómica, ENCB, Instituto Politécnico Nacional, Mexico City, Mexico

[6]Cancer Control Center of the Tachira State, San Cristobal, Venezuela

[7]Grupo de Investigación en Biología del Cáncer, Instituto Nacional de Cancerología, Bogotá, Colombia

[8]Department of Oncology, Wayne State University School of Medicine, Detroit, Michigan

[9]Instituto de Investigaciones en Ciencias de la Salud, National University of Asunción, Asunción, Paraguay

[10]Cancer Institute of Colombia, Bogotá, Colombia

Correspondence
Federico Canzian, Genomic Epidemiology Group, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.
Email: f.canzian@dkfz.de

Abstract

Helicobacter pylori (Hp) infects the stomach of about half of the human population and is strongly associated with the risk of gastric cancer (GC) and its premalignant precursors. The cag pathogenicity island (cagPAI) is a region of the Hp genome encoding for key molecular machinery involved in the infection process. Following a sequencing study, we selected 50 genetic polymorphisms located in seven cagPAI genes and tested their associations with the risk of advanced gastric premalignant lesions and GC in 1220 subjects from various Latin American populations showing the whole spectrum of phenotypes from gastritis to GC. We found that three polymorphisms of *cagA* are associated with the risk of advanced gastric premalignant lesions (incomplete intestinal metaplasia [ie, Type 2 and 3] or dysplasia), and that six polymorphisms located in *cagA*, *cagL* and *cagI* were associated with risk of GC. When corrected for multiple testing none of the associations were statistically significant.

Abbreviations: ASR, age-standardized incidence rate; AUC, area under the curve; cagPAI, cag pathogenicity island; CG, chronic gastritis; CI, confidence interval; CLR, conditional logistic regression; GC, gastric cancer; Hp, *Helicobacter pylori*; IL-8, interleukin 8; IM, intestinal metaplasia; MAF, minor allele frequency; OR, odds ratio; RSCU, relative synonymous codon usage; SIFT, Sorting Intolerant From Tolerant; T4SS, type IV secretion system.

However, scores built by integrating the individual polymorphisms were significantly associated with the risk of advanced gastric premalignant lesions and GC. These results have the potential of establishing markers for risk stratification in the general population, in view of targeting Hp eradication to high-risk population groups.

KEYWORDS

gastric cancer, genetic polymorphisms, *Helicobacter pylori*, pathogenicity island, premalignant gastric lesions

1 | INTRODUCTION

Gastric cancer (GC) is an important public health problem worldwide, as it affects over 1 million people annually, with nearly 800 000 cases dying of the disease in 2018 (https://gco.iarc.fr/today/).[1] Its incidence varies widely among countries, with East Asia and Latin America as high incidence areas (age-standardized incidence rates [ASR] of 22.4 and 8.7, respectively), whereas incidence has been declining steadily in Europe and North America, with current ASR of 8.1 and 6.9, respectively.[1]

The natural history of GC is well known. Acute stomach inflammation can become chronic and lead to atrophic gastritis, which can in turn progress to intestinal metaplasia (IM) and dysplasia, before eventually becoming cancer.[2]

The most important established risk factor for GC is an infection by *Helicobacter pylori* (Hp). Hp is considered a Class 1 carcinogen by the International Agency for Research on Cancer.[3] The association of Hp with GC risk is strong, with odds ratio (OR) in the order of six for noncardia GC.[4] Hp infection has also been reported to be associated with increased risk of premalignant gastric lesions, as well as gastric ulcer.[5] Hp infection is one of the most prevalent infections in the human population. It is estimated that about half of the world population carries Hp,[6] with infection rates in adults ranging from values in the order of 25% in North America to over 60% in Latin America and the Caribbean.[7,8] However, most infected people have only mild (usually in the form of gastritis) or no symptoms at all, and only a small fraction of infected individuals will develop serious sequelae such as advanced premalignant gastric lesions, GC or gastric ulcer.[2,5]

Not all Hp strains carry the same risk of such advanced endpoints. Hp has a very high rate of genetic variability.[9] In particular, the *cagA* pathogenicity island (cagPAI) is a key virulence factor and is crucial in the pathogenesis of Hp-associated diseases.[10] The cagPAI stretches over 40 kb in the Hp genome and encodes for 31 genes that form a type IV secretion system (T4SS) similar to molecular machinery used by other bacteria and consisting of a molecular syringe that injects CagA and other bacterial molecules into cells of the host gastric mucosa.[11] This in turn triggers a cascade of inflammatory events involving production of interleukin 8, inflammation and morphological changes ultimately leading to advanced premalignant gastric lesions and GC.[11]

The *cagA* gene, and indeed the whole cagPAI, may be present or absent in different Hp strains. We demonstrated a strong association of *cagA* presence with increased risk of advanced premalignant gastric lesions and GC.[12] We hypothesized that genetic variability in cagPAI genes at the level of polymorphisms in individual bases might also show association with risk for advanced premalignant gastric lesions and GC. To this end, we performed a first study on a small number of samples from Mexico and Venezuela.[9] Although we identified some polymorphisms associated with GC risk, the study was limited by the small sample size and the lack of samples of subjects with intermediate-risk like IM. We, therefore, launched a new study doing whole-genome sequencing for a total of 75 samples consisting of 37 subjects with chronic gastritis (CG), 21 with IM and 16 with GC cases from Mexico, Colombia and a reference strain. We identified nonsynonymous cagPAI variants that were associated with the risk of IM/GC combined or either lesion.[13] The preliminary results of the sequencing study, however, need replication and validation with a much larger sample size, which we accomplished with the present work.

What's new?

H. pylori is a class 1 carcinogen. However, not all strains increase the risk of gastric cancer. In this study, the authors identified a number of SNPs in the *cagA* pathogenicity island of *H. pylori* that are associated with pre-malignant lesions or gastric cancer. They then developed a scoring system based on these SNPs to quantify an individual patient's level of risk. These markers may enable risk stratification in a population, with the goal of targeting of *H. pylori* eradication to high-risk groups.

2 | MATERIALS AND METHODS

2.1 | Study population

We used samples from multiple studies to represent a wide range of Latin American populations from high- and low-risk of GC, and of the whole spectrum of lesions from gastritis to GC. Details concerning eligibility, recruitment, data and sample collection, endoscopic examination and pathological diagnoses of each study have been published elsewhere.[9,14-20] Subjects were at least 30 years old, and recruited

from four countries in Latin America, with contrasting risk for GC, Venezuela (Tachira province) with high-risk, Colombia with five cities representing both high- and low-risk areas, and two countries with lower risk, Mexico (two hospitals in Mexico City) and Paraguay (two hospitals in Asunción City) in a varied period from the early 1990s to early 2000s; however, in each site recruitment time spanned no more than 3 years and all disease groups were collected within the same time period. All study subjects signed an informed consent and ethical clearance was obtained from the committee of each recruitment as well as of the coordinating center.

Eligible subjects were those who were confirmed positive for *cagA* gene in biopsy samples. From the parent studies, we were able to locate 2114 eligible samples, of sufficient quality from either biopsies or DNA extracted from biopsies (1397 from Venezuela, 325 from Colombia, 262 from Mexico and 130 from Paraguay).

2.2 | Sample preparation

The DNA was extracted from frozen tissues of biopsy samples using QIAamp DNA Micro Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

In cases where biopsy DNA was not available and Hp culture was positive (n = 75 from Mexico and n = 2 from Colombia), DNA of cultured strains was purified using the guanidine thiocyanate-EDTA-Sarkosyl method[21] in Mexico, and with a Pure-Link Genomic DNA Mini Kit (Life Technologies, Carlsbad, California) according to the manufacturer's instructions in Colombia. A whole-genome amplification was carried out on samples with low-levels of DNA (37% of the total), using the Genomiphi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Marlborough, Massachusetts), according to manufacturer's instructions.

2.3 | Polymorphism selection

We aimed to study selected genes of the pathogenicity island cagPAI, namely *cagA*, *cagC*, *cagE*, *cagI* *cagL*, *cagX*, *cagYc* and *cagγ*. They were chosen because their functions have been well characterized in the T4SS, including CagA,[22] CagE[23] and Cagγ,[24] or because they are present extracellularly (ie, as pilus or bacterial surface-related proteins), suggesting possible interactions with host cells, like CagC,[23,25] CagI,[26] CagL,[27] CagX[25] and CagY.[28] Variants on these genes were identified by a whole-genome sequencing study performed on 74 Hp strains isolated from patients from Mexico and Colombia, 37 CG, 21 with IM and 16 with GC.[13] We selected candidate polymorphisms from SNPs with at least 7.5% frequency in the 74 sequenced strains (ie, observed in at least five strains), then applied the following criteria: for non-synonymous variants, OR of ≥1.5 or ≤0.67 for GC + IM compared to CG, and OR ≥ 3.0 or OR ≤ 0.33 for synonymous variants. In addition to these candidate variants, all variants present at the same codon, regardless of their overall frequencies, were included for assay development to ascertain the frequency of candidate variants at a given codon.

To test for samples that are Hp positive but cagPAI negative, we added two polymorphisms in housekeeping genes mapping outside of the cagPAI (atpA185 and ureI60; ie, samples that would give no amplification for markers within the cagPAI but show signals for the two markers in housekeeping genes).

2.4 | Genotyping

We submitted a total of 122 assays for development at Kbiosciences-LGC (Hoddesdon, UK). Triallelic or multiallelic polymorphisms (N = 20) were designed as multiple biallelic assays. For example, two assays were designed for each triallelic polymorphism, each distinguishing the major allele and one of the minor alleles. Complex polymorphisms spanning two or more bases (ie, with two or more polymorphic sites within the same codon resulting in amino acid changes) were broken down into multiple biallelic assays.

Given the very high genetic variability of Hp, not all selected polymorphisms were suitable for the development of genotyping assays. A total of 29 assays failed either at the level of bioinformatic testing, or at validation based on genotyping of up to 96 samples of Hp from cultured strains. Eventually, we obtained 93 validated genotyping assays.

Genotyping was performed at the German Cancer Research Center (DKFZ) in Heidelberg, Germany, in 384-well plates, using KASP technology as recommended by the manufacturer. In addition to the samples, negative controls and duplicated samples (5%) used for quality control purposes were put on each plate and genotyped under the same conditions as the other samples. The personnel performing the genotyping was blind to the diagnosis status. PCR plates were read on a ViiA 7 Real-Time PCR System and genotypes were called with the QuantStudio software (ThermoFisher Applied Biosystems, Waltham, Massachusetts).

2.5 | Quality control and data preprocess

We verified genotype concordance between duplicated samples. This led to the exclusion of 18 samples that had more than two discordant calls. The concordance rate of the duplicates in the remaining samples was 99.25%. We further removed 24 unintended duplicate samples. In addition, samples with a call rate of less than 50% were discarded for subsequent analysis. Finally, we also removed 46 subjects with duodenal ulcer, leaving a total number of 1220 samples for analysis. We checked that allelic frequencies of all polymorphisms were in general agreement with those observed in the sequencing analysis reported in our previous work.[13]

We removed the following assays from the analysis: two polymorphisms with very low minor allele frequency, eight with missing data for more than 33% of the samples; and 52 assays that were individual components of multi-allelic assays, representing 18 polymorphisms, and were replaced by consensus genotypes for a given codon, as well as three assays from multi-allelic polymorphisms, for which the

assay for the other allele could not be developed. The data preprocess step left 52 polymorphisms available for multiple imputations.

2.6 | Statistical and bioinformatic analysis

2.6.1 | Multiple imputation

Multiple imputation imputes incomplete multivariate data by chained equations using fully conditional specification implemented by the MICE R package as described by Van Buuren.[29] Briefly, each variable has its own imputation model. Built-in imputation models are provided based on types of data such as continuous, unordered and ordered categorical data. Each missing value was imputed multiple times. A total number of 30 imputed datasets were created, each with five iterations.

2.6.2 | Univariate association analysis

The outcome of interests was cancer vs noncancer or high-grade vs low-grade premalignant lesions. One by one screening of each marker was performed with conditional logistic regression (CLR) stratified on country and adjusted for baseline covariates, such as age, sex, education status, smoking status, length of refrigerator use and grain intake levels, separately for the two endpoints. R package survival was used on each imputed data set for the CLR model and summarized with R package MICE.

2.6.3 | Multivariable association analysis

Markers that were significant ($P < .05$) from the adjusted univariate CLR model were fit simultaneously into multivariable CLR model. Sensitivity of the multivariable model was also evaluated with additional *P*-value cutoffs, such as $P < .1$ or $P < .2$. An empirical score method was used as well to reduce the dimension of the multivariable CLR model. Briefly, risk alleles of each marker in the multivariable model contributed to the score equally. The score for each sample could range from 0 to the total number of genetic markers in the multivariable CLR from each imputed dataset. The final score for each subject was calculated by averaging through all the 30 imputed datasets. The OR is referred to the risk of each additional risk allele. These analyses were also performed separately for the two endpoints.

2.6.4 | Performance measure

There is no practical approach to assess the area under the curve (AUC) of the CLR model, as the model does not provide predicted risk probability at the level of the individual subject. We provide an alternative AUC for our scoring method. For each subject, the score was scaled to a proportion by dividing it with the total number of genetic markers in the multiple CLR model. This proportion was then directly used as the predicted risk probability, together with the true label of the subject's outcome, to produce the AUC. An AUC value of larger than 0.5 reflects the performance of the score method as a classifier.

The functional effect of nonsynonymous polymorphisms was evaluated with the use of the Sorting Intolerant From Tolerant (SIFT) algorithm (https://sift.bii.a-star.edu.sg).[30] The impact of polymorphisms on codon usage bias was evaluated according to Lafay et al.[31]

3 | RESULTS

The final dataset used for statistical analysis consisted of 1220 samples (Table 1), and included 73.2% of subjects with low-grade premalignant lesions, 14.5% of subjects with high-grade premalignant lesions (consisting of IM Grade 2 and 3, and dysplasia) and 12.3% of subjects with GC.

Fifty polymorphisms located in *cagA* (29 polymorphisms), *cagC* (2), *cagE* (5), *cagI* (4), *cagL* (8) and *cagX* (2) were retained in the final statistical analysis. We performed two sets of analyses. In the first, we compared subjects who had low-grade premalignant lesions (n = 893) with subjects who had high-grade premalignant lesions (n = 177). Four polymorphisms (cagA1283, cagA2551, cagA3490_3491 and cagX31_32) showed associations with increased risk of high-grade lesions. After multivariable association analysis, only cagA1283, cagA2551 and cagA3490_3491 remained significant. Next, we combined the three polymorphisms into a score, whereby the number of risk-increasing alleles at any of the three polymorphisms is added up for each study subject; the score can thus have any value between 0 and 3 (Figure S1). The score was also associated with the risk of high-grade lesions, with an OR = 1.99 for each additional risk allele, and a strong statistical significance ($P = 2.56 \times 10^{-6}$). The AUC of the high-risk lesion score was 0.64 (95% CI 0.60-0.69). Relaxing the threshold for inclusion in the score to $P < .1$ or $P < .2$ resulted in larger numbers of polymorphisms in the scores, but not to significantly better AUC (data not shown). Results of the polymorphisms showing significant associations at $P < .05$ and included in the score are shown in Table 2, whereas results for all polymorphisms are reported in Table S1.

In the second analysis, we compared subjects who did not have GC (n = 1070) with subjects with GC (n = 150). Three polymorphisms (cagA2419, cagA3435 and cagL400) were associated with increased risk of GC and three (cagA1576, cagI1007 and cagL184) were associated with decreased risk. Some pair-wise combinations of the six polymorphisms showed weak but significant correlation (cagL184 and cagL400: $r^2 = .28$, $P = 2.20 \times 10^{-35}$, cagI1007 and cagL184: $r^2 = .19$, $P = 1.82 \times 10^{-19}$, cagI1007 and cagL400: $r^2 = .09$, $P = 2.28 \times 10^{-8}$).

Like for the previous analysis we built a score using the risk-increasing alleles of all six polymorphisms (ie, we considered the reference allele instead of the variant allele for cagA1576, cagI1007 and cagL184). The GC score has values between 0 and 6 (Figure S2). Each additional risk allele in the score gives an OR = 2.42, and this result is strongly significant ($P = 5.41 \times 10^{-8}$). The AUC of the cancer score was 0.65 (95% CI 0.61-0.70). In this case too, using different

TABLE 1 Study population

	Mexico	Paraguay	Colombia	Venezuela	Total
Diagnosis					
Chronic gastritis	84	35	50	271	440
Atrophic gastritis/IM1	22	18	102	311	453
Total low-grade	106	53	152	582	893
IM2 + 3/dysplasia	14	12	40	111	177
Total noncancer	120	65	192	693	1070
Cancer	41	16	15	78	150
Total	161	81	207	771	1220
Male	39.1%	55.6%	50.2%	53.4%	51.5%
Median age (25%-75%)	51 (41-64)	51 (40-62)	52 (42-61)	47 (40-55)	48 (40-58)
Education					
None or less than primary	12.6%	5.2%	15.3%	36.6%	28.2%
Primary	55.6%	66.2%	46.3%	32.4%	39.6%
Secondary or higher	31.8%	28.6%	38.4%	31.0%	32.2%
Tobacco smoking					
Never smokers	60.7%	53.3%	62.3%	66.0%	63.9%
Ex smokers	17.8%	35.1%	28.4%	8.3%	14.6%
Current smokers	21.5%	11.7%	9.3%	25.7%	21.5%
Home refrigerator use, >30 years	28.9%	55.8%	43.0%	38.9%	39.6%
Median grain intake, servings per week (25%-75%)	10 (6-16)	15 (12-23)	18 (14-24)	15 (12-17)	15 (12-18)

Abbreviation: IM, intestinal metaplasia.

TABLE 2 Associations between polymorphisms in Hp cagPAI genes and risk of high-grade premalignant gastric lesions ($P < .05$)

Polymorphism	Major allele DNA	Minor allele DNA	Codon number	Major allele amino acid	Minor allele amino acid	Low-grade Frequency[a]	High-grade frequency[a]	OR[b]	95% CI	P
cagA1283	A	C	428	Asn	Thr	29.9%	41.5%	1.69	1.08-2.64	.021
cagA2551	G	A	851	Ala	Lys	10.1%	19.9%	2.29	1.41-3.71	.001
cagA3490_3491	wild	mut	1162	Cys	Ser	11.7%	23.8%	2.16	1.32-3.54	.002
Score[c]								1.99	1.50,2.63	2.56×10^{-6}

Note: cagX31_32 showed association with risk of high-grade premalignant gastric lesions ($P < .05$) as well (Table S2), but the association was not significant after multivariable analysis, therefore it was not included in the score.

Abbreviation: cagPAI, cag pathogenicity island; HP, *Helicobacter pylori*; OR, odd ratio.

[a]Minor allele frequencies in subjects with low-grade and high-grade premalignant lesions, respectively.

[b]Odds ratios, stratified by country and adjusted for age, sex, education status, smoking status, length of refrigerator use and grain intake levels.

[c]Score composed of the three above polymorphisms. Minor alleles of each marker contributed to the score equally. The score ranges from 0 to 3. The OR is referred to the risk of each additional risk allele.

thresholds for inclusion in the score ($P < .1$ or $P < .2$) did not lead to significantly higher AUC (data not shown). Results of the polymorphisms showing significant associations at $P < .05$ and of the score are shown in Table 3 and the results of all polymorphisms are reported in Table S1.

There were no significant interactions between the six polymorphisms associated with GC risk (15 pair-wise interaction tests, one at a time) or the three polymorphisms associated with risk of high-grade lesions (three pair-wise interaction tests, one at a time; data not shown).

The two polymorphisms in housekeeping genes (atpA185 and ureI60) did not show any association with the risk of either high-grade lesions or GC (data not shown).

GC patients were also subdivided according to the histological classification of their tumors (intestinal vs diffuse). We then performed an exploratory case-case analysis for the six polymorphisms associated with GC risk to see if they could be associated with the risk of one specific histology. None of the polymorphisms showed significantly different frequencies between two histology types (intestinal vs diffuse), with *P*-values ranging from .33 to .87 (Table S3).

TABLE 3 Associations between polymorphisms in Hp cagPAI genes and risk of gastric cancer ($P < .05$)

Polymorphism	Major allele DNA	Minor allele DNA	Codon number	Major allele amino acid	Minor allele amino acid	Noncancer frequency[a]	Cancer frequency[a]	OR[b]	95% CI	P
cagA1576	G	A	526	Ala	Thr	8.1%	2.2%	0.22	0.06-0.78	.020
cagA2419	A	G	807	Arg	Gly	13.3%	18.8%	1.83	1.07-3.13	.028
cagA3435	C	T	1143	Asp	Asp	6.0%	10.7%	2.44	1.16-5.16	.019
cagI1007	C	T	336	Ala	Val	29.4%	24.2%	0.59	0.36-0.97	.039
cagL184[c]	G	C	62	Glu	Lys	11.2%	6.1%	0.46	0.22-1.00	.049
cagL400	A	G	134	Ile	Val	21.2%	30.8%	1.95	1.08-3.54	.28
Score[d]								2.42	1.77-3.30	5.41×10^{-8}

Abbreviation: cagPAI, cag pathogenicity island.

[a]Minor allele frequencies in subjects with and without gastric cancer, respectively.

[b]Odds ratios, stratified by country and adjusted for age, sex, education status, smoking status, length of refrigerator use and grain intake levels.

[c]cagL184 has a third allele, which is not significantly associated with risk of gastric cancer.

[d]Score composed of the six above polymorphisms. Risk alleles of each marker contributed to the score equally. The score ranges from 0 to 6. The OR is referred to the risk of each additional risk allele.

We reported in Figure S3 the positions within the respective genes of the polymorphisms showing associations with the risk of either high-grade premalignant lesions or GC.

All results showed here were generated with data after multiple imputation. Analyses performed using raw data before multiple imputation gave essentially the same results (data not shown).

Study subjects can be subdivided according to their geographic area of origin between regions at low risk of GC, including Mexico, Paraguay and coastal areas of Colombia and regions at high-risk, including Venezuela and mountain areas of Colombia. We compared frequencies of polymorphisms of subjects with low-grade premalignant lesions between low risk (n = 683 subjects) and high-risk regions (n = 210). None of the three polymorphisms associated with the risk of high-grade premalignant lesions or the six polymorphisms associated with the risk of GC showed significant differences between the regions (data not shown).

None of the associations we reported for the individual polymorphisms are significant if multiple testing is taken into account (with a Bonferroni-corrected threshold of 0.05/(53 polymorphisms [50 polymorphisms, three of which are triallelic] ×2 sets of analyses (low-grade vs high-grade and noncancer vs cancer)) = 0.00047). Evaluation of the results with a false discovery rate also shows that none of the individual associations are significant (data not shown).

None of the nonsynonymous variants associated with risk of either high-grade lesions or GC was predicted by SIFT to have a relevant functional effect, with the exception of cagL184, where the polymorphism we found to be associated with risk of GC causes an amino acid substitution, with the major allele coding for glutamine and the minor allele coding for lysine. This replacement is predicted to be not tolerated and possibly resulting in disruption of the protein function.

We also analyzed the relative codon usage according to Lafay et al.[31] Both alleles of polymorphism cagA3435, which is associated with the risk of GC, code for aspartic acid. However, the major allele has a relative synonymous codon usage (RSCU) of 0.54, while the minor allele has a value of 1.46. The RSCU is the observed frequency of a codon, divided by the frequency expected if all possible codons for that amino acid were used equally. Thus, RSCU = 1 indicates a lack of bias, RSCU < 1 shows that a codon is underrepresented and RSCU > 1 that a codon is overrepresented with respect to the expected.

4 | DISCUSSION

We studied the possible associations between polymorphisms in key genes of the Hp cagPAI and risk of high-grade premalignant gastric lesions and of GC in several Latin American populations. Our study is the first to examine the associations of HP cagPAI sequence variants with gastric pathology in unselected gastric biopsy specimens, using high throughput genotyping assays. To date, Hp genotyping on biopsy specimens has been almost limited to CagA EPIYA motif patterns.[32] There are more recent studies based on sequencing technology, reporting associations between specific Hp genetic variants and clinical phenotypes.[33-37] However, these sequencing studies have been performed on a limited number of samples and with Hp strains isolated from the patients. It is important to note that Hp culture from clinical specimens is not always successful. Even if Hp is detectable by PCR or by histology, culture often fails, indicating potential serious bias in the studies based only on cultured strains. In addition, such strain-based studies often lack covariate information relevant to GC, which is very relevant considering that GC is a multifactorial disease.

In our previous study with Hp from patients with CG, IM or GC from Mexico and Colombia,[13] a large number of genetic polymorphisms were found in the cagPAI genes studied. In the present work, we decided to analyze polymorphisms with increased possibility of being functionally relevant and study their possible associations with risk for high-grade premalignant gastric lesions and for GC. We started from a large pool of candidate polymorphisms, but reduced it

to 50 polymorphisms, representing a fraction of our initial list. Development of genotyping assays was one of the major difficulties we faced in this project. The very high-genetic diversity of Hp, particularly in the cagPAI, makes it prohibitive to develop genotyping assays for many polymorphisms. As a result, we lost many candidate SNP located in hypervariable regions, leaving more SNP located in rather conserved regions, which may in fact be relevant for gene function.

To ensure a sufficient level of quality we discarded samples with a call rate lower than 50%, resulting in a loss of over 40% of the initial samples. Consequently, to cope with the high rate of missing data in the remaining dataset we chose to apply multiple imputations. We compared the results obtained from data before and after multiple imputation and we did not observe notable differences, which suggests that multiple imputation performed reliably and that our final results are robust.

We found that three polymorphisms in *cagA* were associated with risk of high-grade premalignant gastric lesions and six polymorphisms in *cagA*, *cagI* and *cagL* were independently associated with risk of GC.

Once cagA is internalized into gastric epithelial cells it interacts with a myriad of intracellular targets including kinases or proteins of the cytoskeleton. Two of the positions we found associated with the disease are located in Domain II (cagA1283 associated with high-grade preneoplasia and cagA1576 with GC) which is a region important for the interaction of cagA with the inner left of the cytoplasmic membrane. Another position is located in Domain III (cagA2419, associated with GC risk) that may affect the intramolecular N-terminal/C-terminal interaction important for the recruitment of PAR1,[38] which in turn may lead to the activation of the Ras-ERK MAPK pathway.

Until recently, the main focus of CagL variants has been hypervariable-amino acid residues 58 to 62[39] located upstream of the RGD motif that is crucial for host integrin binding,[40] and the cagL184 polymorphism found in our study resides at this region (residue 62). In accordance with this, a meta-analysis of worldwide strains[39] found an overall positive association between polymorphisms in the 58 to 62 region and GC. Yeh et al showed that polymorphisms at residues 58 and 59 induce a corpus shift of gastric integrin $\alpha5\beta1$,[33] although an in vitro study in AGS cells that tested various combinations of amino acids at residues 58 and 59 did not find functional differences.[41] We genotyped cagL172 (codon 58) and did not find any significant association. However, our analysis with SIFT predicts that the amino acid substitution caused by cagL184 at codon 62 is not tolerated and likely to disrupt the function of the protein. As noted by Tafreshi et al,[41] it is possible that variation at residues 58 and 59 might work in concert with variation at residues 60 to 62 in influencing the structural integrity of CagL and therefore its function.[41] More recent sequencing studies have identified a number of variants in this gene. The other *cagL* variant associated with GC in our study, on residue 134, has been reported with variable results in other studies in Latin America[34,36,37]; although this variant has not been detected in East Asian strains.[33,35] Thus, although the data from these Western strains are inconclusive in terms of its association with GC, this *cagL* variant is likely to represent a Western strain-specific marker. Furthermore, residue 134 is located within a structurally important region of *cagL* containing a disulfide bond that bridges helix 5 to the C-terminus of the short $\alpha4$ helix.[42,43]

One study that examined sequence variants of *cagI*, in addition to those of *cagL*, of East Asian strains, did not detect the cagI1007 variant at residue 336 that showed the association with GC in our study.[35]

When we took into account multiple testing, none of the associations of the individual polymorphisms remained significant; however, scores generated by the combinations of the individually associated polymorphisms were associated with the risk of high-grade premalignant gastric lesions and of GC, respectively, with strong statistical significances. Taken together, these results suggest that individual variants, if confirmed in further studies, could be useful to gain insight about molecular mechanisms of gastric carcinogenesis, but not likely to be of much use for risk stratification in the population. On the other hand, scores could be useful for risk stratification in the general population, in view of targeting Hp eradication in population groups at particularly high-risk of developing high-grade premalignant lesions or GC. The AUC we obtained for both scores were promising, although clearly not at the level where it can be envisaged to use these scores as predictive tools in the general population. However, it is likely that additional studies of genetic variants located in other Hp genes will uncover further risk-associated polymorphisms. Future versions of the scores generated with the larger number of risk variants will have better predictive power.

Exploratory analyses of our previous discovery work[13] found 19 polymorphisms showing associations with the risk of IM and/or GC ($P < .05$). Six of them could be studied here, however, none of them showed significant associations with either IM or GC risk. Major reasons for these differences may include the much smaller sample size of our previous analysis (74 samples in total), the inclusion of samples from additional different countries in the current study and differences between culturable and unculturable strains. Additional minor factors may include the fact that diagnoses were grouped in different ways, with the high-grade premalignant gastric lesions consisting of only IM cases in Rizzato et al, and of IM2 + IM3 + dysplasia here and with or without adjustment of covariates.

The main strengths of our work are the inclusion of unculturable Hp strains and a large sample size, even taking into account the loss of many subjects due to the low quality of the samples. Sample size remains relatively large when considering the different subgroups of subjects, with 177 subjects with high-grade premalignant gastric lesions and 150 GC cases.

Limitations include the fact that we studied only Latin American populations and therefore findings are not necessarily generalizable to other populations, considering the known large differences between Hp strains of different populations.[44,45] On the other hand, Latin America is one of the areas with the highest Hp infection rates and with the highest GC mortality rates.[1] Hispanic populations also represent large minorities in the USA and other countries. Additionally, this work did not include large structural variations. However, we studied EPIYA and CM motifs in our sequencing study and we did not observe any association with risk of IM and/or GC.[13] Moreover, structural

variations are meaningfully studied only by sequencing, which most of our samples are unsuitable for. Another limitation is the lack of data on subjects without any stomach pathology. However, collecting gastric endoscopies from totally asymptomatic people is not acceptable on ethical grounds, and people with symptoms justifying a stomach endoscopy usually have at least gastritis. In particular, in Latin America Hp infection is present in over 70% of the adult population[46] and infection invariably causes inflammation of the gastric mucosa, usually asymptomatic; thus, most general population has Hp infection and gastritis.

We chose KASP assays, which are based on competitive allele-specific PCR, suitable for low quality DNA. However, the call rates of our genotyping data were lower than we anticipated. We speculate the primary reason for the low call rates is quantity of Hp DNA, rather than DNA degradation over time. Gastric biopsies primarily are comprised of human cells, and even in Hp-positive patients, bacterial cell counts are minuscule compared to human cells, while DNA content is several orders of magnitude lower in bacterial cells than in human cells. Although we employed WGA to overcome this limitation, Monstein et al report poorer performance of WGA for bacterial DNA in substantial excess of human DNA.[47] An additional contributing factor to the low call rates may be the presence of unidentified sequence deviations that prevented amplification with designed primers and probes, given the high variability of these genes. For alternative approaches, such as sequencing the whole region of interest, an additional step of bacterial DNA enrichment or mammalian DNA depletion may be warranted.

In conclusion, we performed a large-scale analysis of Hp cagPAI polymorphisms in relation to risk of high-grade premalignant gastric lesions and GC. We found a few associations of individual polymorphisms, which did not reach statistical significance when multiple testing was taken into account. However, when polymorphisms were combined in scores, they showed strong associations with the risk of both high-grade premalignant gastric lesions and of GC. These results have the potential of establishing markers for risk stratification in the general population, in view of targeting Hp eradication to high-risk population groups.

ACKNOWLEDGEMENTS

This work was supported by the National Cancer Institute, National Institutes of Health, United States (5R21CA182822, P.I. I. K.).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA ACCESSIBILITY

The primary data for this work will be made available to researchers who submit a reasonable request to the corresponding author, conditional to approval by all the collaborators. Data will be stripped from all information allowing identification of study participants. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

All study subjects signed an informed consent and ethical clearance was obtained from the committee of each recruitment as well as of the coordinating center.

ORCID

Federico Canzian https://orcid.org/0000-0002-4261-4583

REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019;144:1941-1953.
2. Correa P, Piazuelo MB. Natural history of *Helicobacter pylori* infection. *Dig Liver Dis*. 2008;40:490-496.
3. International Agency for Research on Cancer. Infection with Helicobacter pylori. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Lyon, France: International Agency for Research on Cancer; 1994:177-240.
4. Webb PM, Law M, Varghese C, Forman D. Gastric cancer and *Helicobacter pylori*: a combined analysis of 12 case control studies nested within prospective cohorts. *Gut*. 2001;49:347-353.
5. Uemura N, Okamoto S, Yamamoto S, et al. *Helicobacter pylori* infection and the development of gastric cancer. *N Engl J Med*. 2001;345:784-789.
6. Parkin DM. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer*. 2006;118:3030-3044.
7. Peleteiro B, Bastos A, Ferro A, Lunet N. Prevalence of *Helicobacter pylori* infection worldwide: a systematic review of studies with national coverage. *Dig Dis Sci*. 2014;59:1698-1709.
8. Zamani M, Ebrahimtabar F, Zamani V, et al. Systematic review with meta-analysis: the worldwide prevalence of *Helicobacter pylori* infection. *Aliment Pharmacol Ther*. 2018;47:868-876.
9. Rizzato C, Torres J, Plummer M, et al. Variations in *Helicobacter pylori* cytotoxin-associated genes and their influence in progression to gastric cancer: implications for prevention. *PLoS One*. 2012;7:e29605.
10. Censini S, Lange C, Xiang Z, et al. Cag, a pathogenicity Island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc Natl Acad Sci U S A*. 1996;93:14648-14653.
11. Hatakeyama M. *Helicobacter pylori* CagA—a bacterial intruder conspiring gastric carcinogenesis. *Int J Cancer*. 2006;119:1217-1223.
12. Plummer M, van Doorn LJ, Franceschi S, et al. *Helicobacter pylori* cytotoxin-associated genotype and gastric precancerous lesions. *J Natl Cancer Inst*. 2007;99:1328-1334.
13. Rizzato C, Torres J, Obazee O, et al. Variations in cag pathogenicity island genes of *Helicobacter pylori* from Latin American groups may influence neoplastic progression to gastric cancer. *Sci Rep*. 2020;10:6570.
14. Muñoz N, Kato I, Peraza S, et al. Prevalence of precancerous lesions of the stomach in Venezuela. *Cancer Epidemiol Biomarkers Prev*. 1996;5:41-46.
15. Kato I, Vivas J, Plummer M, et al. Environmental factors in *Helicobacter pylori*-related gastric precancerous lesions in Venezuela. *Cancer Epidemiol Biomarkers Prev*. 2004;13:468-476.
16. Reyes-Leon A, Atherton JC, Argent RH, Puente JL, Torres J. Heterogeneity in the activity of Mexican *Helicobacter pylori* strains in gastric epithelial cells and its association with diversity in the cagA gene. *Infect Immun*. 2007;75:3445-3454.
17. de la Trejo OA, Torres J, Pérez-Rodríguez M, et al. TLR4 single-nucleotide polymorphisms alter mucosal cytokine and chemokine patterns in Mexican patients with *Helicobacter pylori*-associated gastroduodenal diseases. *Clin Immunol*. 2008;129:333-340.
18. Martínez T, Hernández-Suárez G, Bravo MM, et al. Association of interleukin-1 genetic polymorphism and CagA positive *Helicobacter pylori* with gastric cancer in Colombia. *Rev Med Chil*. 2011;139:1313-1321.
19. Flores-Luna L, Camorlinga-Ponce M, Hernandez-Suarez G, et al. The utility of serologic tests as biomarkers for *Helicobacter pylori*-associated precancerous lesions and gastric cancer varies between Latin American countries. *Cancer Causes Control*. 2013;24:241-248.

20. Trejo-De La OA, Torres J, Sánchez-Zauco N, et al. Polymorphisms in TLR9 but not in TLR5 increase the risk for duodenal ulcer and alter cytokine expression in the gastric mucosa. *Innate Immun*. 2015;21:706-713.
21. Clabots CR, Johnson S, Bettin KM, et al. Development of a rapid and efficient restriction endonuclease analysis typing system for clostridium difficile and correlation with other typing systems. *J Clin Microbiol*. 1993;31:1870-1875.
22. Nishikawa H, Hatakeyama M. Sequence polymorphism and intrinsic structural disorder as related to pathobiological performance of the *Helicobacter pylori* CagA oncoprotein. *Toxins*. 2017;9:136.
23. Kerr JE, Christie PJ. Evidence for VirB4-mediated dislocation of membrane-integrated VirB2 pilin during biogenesis of the agrobacterium VirB/VirD4 type IV secretion system. *J Bacteriol*. 2010;192:4923-4934.
24. Zhong Q, Shao S, Mu R, et al. Characterization of peptidoglycan hydrolase in cag pathogenicity Island of *Helicobacter pylori*. *Mol Biol Rep*. 2011;38:503-509.
25. Andrzejewska J, Lee SK, Olbermann P, et al. Characterization of the pilin ortholog of the *Helicobacter pylori* type IV cag pathogenicity apparatus, a surface-associated protein expressed during infection. *J Bacteriol*. 2006;188:5865-5877.
26. Kumar N, Shariq M, Kumari R, Tyagi RK, Mukhopadhyay G. Cag type IV secretion system: CagI independent bacterial surface localization of CagA. *PLoS One*. 2013;8:e74620.
27. Wiedemann T, Hofbaur S, Tegtmeyer N, et al. *Helicobacter pylori* CagL dependent induction of gastrin expression via a novel αvβ5-integrineintegrin linked kinase signalling complex. *Gut*. 2012;61:986-996.
28. Barrozo RM, Cooke CL, Hansen LM, et al. Functional plasticity in the type IV secretion system of *Helicobacter pylori*. *PLoS Pathog*. 2013;9:e1003189.
29. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16:219-242.
30. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40:W452-W457.
31. Lafay B, Atherton JC, Sharp PM. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*. 2000;146:851-860.
32. Li Q, Liu J, Gong Y, Yuan Y. Association of CagA EPIYA-D or EPIYA-C phosphorylation sites with peptic ulcer and gastric cancer risks: a meta-analysis. *Medicine (Baltimore)*. 2017;96:e6620.
33. Yeh YC, Chang WL, Yang HB, Cheng HC, Wu JJ, Sheu BS. *H. pylori* cagL amino acid sequence polymorphism Y58E59 induces a corpus shift of gastric integrin α5β1 related with gastric carcinogenesis. *Mol Carcinog*. 2011;50:751-759.
34. Cherati MR, Shokri-Shirvani J, Karkhah A, Rajabnia R, Nouri HR. *Helicobacter pylori* cagL amino acid polymorphism D58E59 pave the way toward peptic ulcer disease while N58E59 is associated with gastric cancer in north of Iran. *Microb Pathog*. 2017;107:413-418.
35. Ogawa H, Iwamoto A, Tanahashi T, et al. Genetic variants of *Helicobacter pylori* type IV secretion system components CagL and CagI and their association with clinical outcomes. *Gut Pathog*. 2017;9:21.
36. Román-Román A, Martínez-Santos VI, Castañón-Sánchez CA, et al. CagL polymorphisms D58/K59 are predominant in *Helicobacter pylori* strains isolated from Mexican patients with chronic gastritis. *Gut Pathog*. 2019;11:5.
37. Yadegar A, Mohabati Mobarez A, Zali MR. Genetic diversity and amino acid sequence polymorphism in *Helicobacter pylori* CagL hypervariable motif and its association with virulence markers and gastroduodenal diseases. *Cancer Med*. 2019;8:1619-1632.
38. Hayashi T, Senda M, Morohashi H, et al. Tertiary structure-function analysis reveals the pathogenic signaling potentiation mechanism of *Helicobacter pylori* oncogenic effector CagA. *Cell Host Microbe*. 2012;12:20-33.
39. Gorrell RJ, Zwickel N, Reynolds J, Bulach D, Kwok T. *Helicobacter pylori* CagL hypervariable motif: a global analysis of geographical diversity and association with gastric cancer. *J Infect Dis*. 2016;213:1927-1931.
40. Bonsor DA, Pham KT, Beadenkopf R, et al. Integrin engagement by the helical RGD motif of the *Helicobacter pylori* CagL protein is regulated by pH-induced displacement of a neighboring helix. *J Biol Chem*. 2015;290:12929-12940.
41. Tafreshi M, Zwickel N, Gorrell RJ, Kwok T. Preservation of *Helicobacter pylori* CagA translocation and host cell proinflammatory responses in the face of CagL hypervariability at amino acid residues 58/59. *PLoS One*. 2015;10:e0133531.
42. Barden S, Schomburg B, Conradi J, Backert S, Sewald N, Niemann HH. Structure of a three-dimensional domain-swapped dimer of the *Helicobacter pylori* type IV secretion system pilus protein CagL. *Acta Crystallogr D Biol Crystallogr*. 2014;70:1391-1400.
43. Choi JM, Choi YH, Sudhanva MS, et al. Crystal structure of CagL from *Helicobacter pylori* K74 strain. *Biochem Biophys Res Commun*. 2015;460:964-970.
44. Thorell K, Yahara K, Berthenet E, et al. Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. *PLoS Genet*. 2017;13:e1006546.
45. Muñoz-Ramírez ZY, Mendez-Tenorio A, Kato I, et al. Whole genome sequence and phylogenetic analysis show *Helicobacter pylori* strains from Latin America have followed a unique evolution pathway. *Front Cell Infect Microbiol*. 2017;7:50.
46. Porras C, Nodora J, Sexton R, et al. Epidemiology of *Helicobacter pylori* infection in six Latin American countries (SWOG trial S0701). *Cancer Causes Control*. 2013;24:209-215.
47. Monstein HJ, Olsson C, Nilsson I, Grahn N, Benoni C, Ahrné S. Multiple displacement amplification of DNA from human colon and rectum biopsies: bacterial profiling and identification of *Helicobacter pylori*-DNA by means of 16S rDNA-based TTGE and pyrosequencing analysis. *J Microbiol Methods*. 2005;63:239-247.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Canzian F, Rizzato C, Obazee O, et al. Genetic polymorphisms in the cag pathogenicity island of *Helicobacter pylori* and risk of stomach cancer and high-grade premalignant gastric lesions. *Int. J. Cancer*. 2020;1–9. https://doi.org/10.1002/ijc.33032