Detection and variability analyses of CRISPR-like loci in the *H. pylori* genome

Jerson Alexander Garcia-Zea Corresp., [1] , **Roberto de la Herrán** [1] , **Francisca Robles Rodríguez** [1] , **Rafael Navajas-Pérez** [1] , **Carmelo Ruiz Rejón** [1]

[1] Departamento de Genética, Universidad de Granada, Granada, Spain

Corresponding Author: Jerson Alexander Garcia-Zea
Email address: alexander7719@correo.ugr.es

Helicobacter pylori is a human pathogenic bacterium with a high genomic plasticity. Although the functional CRISPR-Cas system has not been found in its genome, CRISPR like loci have been recently identified. In this work, 53 genomes from different geographical areas are analyzed for the search and analysis of variability of this type of structure. We confirm the presence of a locus that was previously described in the VlpC gene in al lgenomes, and we characterize new CRISPR-like loci in other genomic locations. By studying the variability and gene location of these loci, the evolution and the possible roles of these sequences are discussed. Additionally, the usefulness of this type of sequences as a phylogenetic marker has been demonstrated, associating the different strains by geographical area.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27196v1 | CC BY 4.0 Open Access | rec: 12 Sep 2018, publ: 12 Sep 2018

Detection and variability analyses of CRISPR-like loci in the *H. pylori* genome

CRISPR-like in *H. pylori*

Jerson Alexander García-Zea [1], Roberto de la Herrán [2], Francisca Robles Rodríguez [3], Rafael Navajas-Pérez [4], Carmelo Ruiz Rejón [1].

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain.

[2] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain.

[3] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain.

[4] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain.

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain.

Corresponding author:

Alexander García [1]

Avda. Fuentenueva s/n, 18071 Granada, Spain. Tel. +34 958-24-30-80. Fax. +34 958-24-40-73.

Email address: alexander7719@correo.ugr.es

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27196v1 | CC BY 4.0 Open Access | rec: 12 Sep 2018, publ: 12 Sep 2018

Abstract

Helicobacter pylori is a human pathogenic bacterium with a high genomic plasticity. Although the functional CRISPR-Cas system has not been found in its genome, CRISPR-like loci have been recently identified. In this work, 53 genomes from different geographical areas are analyzed for the search and analysis of variability of this type of structure. We confirm the presence of a locus that was previously described in the VlpC gene in all genomes, and we characterize new CRISPR-like loci in other genomic locations. By studying the variability and gene location of these loci, the evolution and the possible roles of these sequences are discussed. Additionally, the usefulness of this type of sequences as a phylogenetic marker has been demonstrated, associating the different strains by geographical area.

Subjects Bioinformatics, Evolutionary Studies, Genetics, Genomics, Microbiology

Key words *Helicobacter pylori*, variability CRISPR-like, *VlpC* gene, phylogenetic marker.

Introduction

The genus *Helicobacter* comprises 20 formally validated species. Within these group, *H. pylori* is worth noting due to its characteristics, as it can be considered as a model organism for the study of genetics and evolution. *H. pylori* has a great genomic plasticity, presenting high rates of mutation and recombination that allows for the generation of new alleles, allowing it to adapt to relatively specific and well-defined habitats such as the stomach and the duodenum [1]. It is well established that *H. pylori* is a highly competent bacterium, and different strains can be found living together in the gastric environment, bringing the populations of *H. pylori* closer to panmixia [2-3]. Genome comparative analyses from diverse origins have shown that this bacterium shows a high degree of genetic diversity, ranging from nucleotide polymorphisms to genetic mosaicism [4].

The CRISPR-Cas system is a defense mechanism against foreign genetic elements derived from bacteriophages, plasmids or extracellular chromosomal DNA [5-6].The CRISPR-Cas loci are variable in number between bacteria and strains [7], and its typical structure is characterized by a CRISPR matrix, a nearby Cas-gene locus, and an AT-rich leader region [8]. This system is also characterized by its rapid evolution and variability which makes its classification a highly complex task, due to the frequent modular recombination of the CRISPR [9] matrix, which may mean that not all CRISPR systems carry the same components [7-10] or fulfill the same functions [6].

The CRISPR-Cas systems have been identified in approximately 40% of the bacteria and 90% of the archaea. However, Burstein et al., 2016 [11] recently proposed that CRISPR-Cas systems are present in only 10% of the archaea and bacteria. This difference in the presence of the CRISPR-Cas system in prokaryotes could due to the fact that the system may not exist in the main non-cultivable bacterial lineages and in those whose lifestyle was symbiotic [11-12].

In the genus *Helicobacter*, the CRISPR-Cas system has only been detected in *H. cinaedi* and *H. mustelae* [13-14], both pathogenic species, but not in H. pylori. However, Bangpanwimon et al., 2017 [15] have more recently described CRISPR-like sequences in the genome of *H. pylori*, more precisely located in the VacA-like paralogue gene (VlpC, HP0922), that could be related to the ability to colonize the stomach [16], suggesting that they could have a regulatory role [17]. In fact, in recent years, hypotheses involving CRISPR loci in the regulation of genes, a function analogous to the functions of RNAi in eukaryotes [18], have appeared where the CRISPR spacers coincided with genes from the genome itself, with important cellular functions (housekeeping) [19]. Another relationship established between CRISPR and pathogenicity has been discussed in strains of *E. coli* and other species, where the interference of CRISPR prevented the acquisition of virulence genes [20]. On the other hand, a reduced content of CRISPR repeats has also been correlated with a greater likelihood that a strain exerts pathogenicity (potential ability to cause disease) [20]. All of these data exemplify the versatility of CRISPR-Cas systems and suggest roles beyond canonical interference against strange genetic elements [21]. The presence of CRISPR orphans of non-vestigial subtype I-F and E in *E. coli* (CRISPR without cas-genes) have been attributed to a possible habitat change, where their presence would be counterproductive [22], granting them a regulatory role, whose spacers could prevent the acquisition of cas (anti-cas) genes, thus facilitating the acquisition of genetic material and increasing biological aptitude [22].

In this work, we analyze the presence and variability of CRISPR-like sequences in *H. pylori* by studying 53 strains, finding that there are several CRISPR-like sequences in their genomes, which are relatively conserved among strains and can be grouped by geographic area. We discuss their possible role in the generation of variability as well as in the regulation of the genes into which they are inserted.

Materials and methods

Analysis of CRSIPR-like loci in *Helicobacter pylori* the sequences of 53 complete genomes **(Table 1)** (GenBank and fasta formats) of different *H. pylori* strains were downloaded from the genomic resource database of the National Biotechnology Information Center [23] (ftp: //ftp.ncbi.nlm.nih.gov/genomes/). To characterize the CRISPR region in the *H. pylori* genomes we used the CRISPRFinder program with default parameters [24] (http://crispr.i2bc.paris-saclay.fr/Server/). In addition, to characterize the CRISPR-like regions in all the genomes analyzed in this work, multiple alignments were created with the Muscle program [25].

We used CRISPRsBlast (E-value: 0.01) to determine the similarity between the direct repeats (DRs) and spacers of the CRISPR loci detected in *H. pylori* and the sequences of DRs and confirmed spacers deposited in the BLAST CRISPR database (http://crispr.i2bc.paris-saclay.fr/crispr/BLAST/CRISPRsBlast.php).

The spacers were also blasted with default parameters against the CRISPRTarget server, which predicts the most likely targets of the CRISPR RNAs (http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html) [26]. The databases used were: mobile genetic elements and phages, viruses.

For phylogenetic analyses we used the Mega7 program [27] with the following parameters: 1000 bootstrap method and Jukes Cantor model.

Identification of operons linked to CRISPR-like and Cas domains

The research on operons linked to the CRISPR-like structure was carried out using the OperonDB database [28] (http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi)

For the identification of cas domains, the HMMs profiles (Markov Hidden Models Profile) of the Cas families were downloaded from TIGRFAM (ftp://ftp.jcvi.org/pub/data/TIGRFAMs/) as well as the Cas proteins described by Haft et al., 2005 [29]. The search of cas proteins was carried out with HMMER software v3.1b2 30, implementing the option 'hmmscan' (search in proteins against collections of proteins of the 53 genomes), with an E-value 10e-5.

Identification of Vac-like gene (vlpC)

To identify and determine the presence of vacA-like gene (VlpC) in the 53 genomes of *H. pylori* used in this study, the reference sequence of strain J99 were downloaded from NCBI: WP_000874591.1 (VlpC). This sequence was blasted against the 53 *H. pylori* genomes with the following parameters: E- value: 10e-5, query coverage >75%.

Also, to determine if the corresponding mRNA of the VlpC gene of the different strains of *H. pylori* was expressed, the cDNA sequences of the 53 genomes were downloaded via FTP (http://bacteria.ensembl.org/info/website/ftp/ index.html) and used as a target to be blasted with the CRISPR-like sequences detected in the VlpC gene, using an E-value of 10e-5.

In addition, for genes that showed a CRISPR-like sequence outside of VlpC, their presence in all genomes was verified using blastn with Geneious v 6.1.8 [31] using an E-value of 10e-5.

To determine genomic rearrangements and possible break-point involved in recombination events, the Mauve software was used for complete alignment of genomes [32].

Results

CRISPR-like loci identification

A total of 53 *H. pylori* assembled and annotated genomes from different geographical regions were analyzed with CRISPRFinder software. Twenty-two CRISPR-like loci were found in 20 strains, with 19 of them exhibiting one CRISPR-like locus and only one strain, SJM180 **(Table 1)** showing three CRISPR-like loci. Of all loci, 16 were located within a VacA-like gene (VlpC gene), with four DRs and three spacer sequences. This gene was integrated in an operon with the genes *OMP*, *4-oxalocrotonate tautomerase*, *recR, truD, htpX, folE, IspA* and, *surE*, in this order [24]. The remaining six CRISPR-like loci were present in other locations of the genome. More specifically, they were located in: a) the BM012A (Australian origin) and Shi470 (Peru origin) strains in a Poly E-rich gene rich protein; b) the Shi417 and Shi112 (both Peru origin) strains within a hypothetical protein (with GO term COG119), and; c) the SJM180 (Peru origin) strain, with two additional loci, with these located in two different hypothetical protein genes (Table 1). For these 22 loci, which were detected with CRISPRFinder, 95 direct repeat sequences (DRs) were identified, being present in 4 to 7 sequences per CRISPR-like locus and ranging from 23 to 36bp in length **(Table S1)**. No similarities were found when these sequences were blasted against the CRISPRsBlast database. A total of 73 spacers were detected ranging in number from 3 to 6 sequences per locus, with lengths ranging between 16 to 69bp. Using CRISPRTarget software, 5 spacers showed similarities to phage, plasmids or viruses sequences **(Table S2)**.

Consensus DRs for each locus and spacers were used to carry out a phylogenetic study. For DRs, two main groups were observed in the phylogenetic tree: one, including the DRs of the six CRISPR-like loci located out of VlpC gene, and the other, with the strains that had the loci within the VplC gene **(Fig. 1)**. The spacers in the phylogenetic tree could be divided into four main groups: three of them corresponded to the group of spacers present in the first, second and third position within the CRISPR-like loci located within the VlpC gene. The fourth group corresponded to the spacers of the CRISPR-like loci found in other genes within the SJM180 (CRISPR1-like and CRISPR3-like), BM012A, Shi417 and Shi112 strains **(Fig. 2)**.

Analysis of CRISPR-like sequences located within VlpC

The VlpC gene was present in all genomes, except for strain Aklavik86. A manual construction of multiple alignments allowed us to determine the presence of a CRISPR-like structure within the VlpC gene for all genomes. Only the South Africa20 strain showed the VlpC gene but not the CRISPR-like locus, as the gene is truncated in the 5' region where this structure would be found. The CRISPR-like locus possessed different degrees of variability between strains. The alignment allowed for an in-depth study of DRs and spacers for this locus. It was observed that the variation of the CRISPR-like structure in the VplC gene was mainly due to the complete duplication and/or deletion of spacers and DRs **(Fig. S1)**. The sequences from the 51 CRISPR-like loci detected in VlpC were used to carry out a phylogenetic analysis. Three clusters were observed, created by grouping the sequences according to their geographical origins **(Fig. 3)**. The first group included the Africa and Europe strains (group A), the second included the Asia (group B) strains and with the last being the Amerindian strains (group C).

Despite the great variability detected, when the transcriptomes of the *H. pylori* strains were analyzed, it was found that the gene corresponding to VlpC mRNA was expressed in 50 of the 52 genomes that possessed this gene, including the CRISPR-like sequence **(Table S3)**.

When a blastn (E-value: 10e-5, query coverage> 75%) was performed using the VlpC gene sequence from *H. pylori* against the genomes of other *Helicobacter* species, only *H. cetorum* showed the presence of this gene. This gene had the CRISPR-like structure, similar to H. pylori and an identity above 80% in DRs, indicating that it was the same locus.

Analysis of CRISPR-like sequences located outside the VlpC gene

In addition to the 16 CRISPR-like loci detected in the VlpC gene by CRISPRFinder, we detected two additional loci in the Shi417 and Shi112 (WP_000536430 and Shi112 WP_000536429 hypothetical protein, respectively) strains, which had identical sequences in their DRs and spacers. These were located in the 5' region of a gene from a hypothetical protein, between the positions 55,000 to 56,000 of the genome. The ontology analysis showed that this protein had domains related to cell division and cycle control. The CRISPR-like locus of this gene had a length of 150bp, with four 23bp DRs and three 19bp spacers of. No similarities were found with other types of genetic element.

When blastn (E-value: 10e-5, query coverage > 75%) was performed using the sequence of this gene against the remaining 51 genomes, it was found in 12 more strains. The 5´ regions were low conserved, and even three strains (aklavik86, aklavik117 and P12) had this region truncated. Whereas the 3' region was highly conserved (85%) for the twelve strains **(Fig. S2)**. All the genes had a CRISPR-like locus in their sequence but, as the CRISPR-like locus is located in the 5' region of the genes, they were degenerate (56% of identity). The origins of these 14 strains with CRISPR-like locus were Amerindian (6) European (6) and African (2), and the phylogenetic tree constructed, using the CRISPR-like sequences, clearly separated these three groups **(Fig. S3)**.

The Shi470 and BM012A strains showed a CRISPR-like locus within a Poly-E rich protein gene (WP_00078209, WP_023591955 respectively). In Shi470, this gene was located between the positions 320.726 and 322.187. In the case of BM012A, it was between the positions 659.636 and 661.240. In a genomic structural analysis of these two strains with Mauve software [32], it was verified that it was the same gene present in a syntenic region but affected by a genomic rearrangement. This gene was included in an inverted segment and near a breakpoint **(Fig. S4a).** The alignment of this gene from both strains revealed a middle location of the CRIPR-like locus, with a 70% similarity. The divergences could be explained for the different number of DRs and spacers detected among them **(Fig. S4b)**.The CRISPR-like locus of Shi470 had a length of 660bp with seven DRs and six spacers while that for BM012A was 174bp in length with four DRs and three spacers. It was interesting to note that the spacers of the Shi470 strain showed similarity with mobile elements and phages **(Table S2)**, while no similarities were found for those from the BM012A strain.

A Blastn (previous parameters) search with the rest of the genomes (51) allowed us to identify this gene in 33 more strains, all of them with CRISPR-like features. The genes showed a high identity (close to 80%) in their sequence except in the CRISPR-like region (60% identity) **(Fig. S5)**. The phylogenetic tree, constructed with the CRISPR-like sequences, clearly separates the four geographic regions **(Fig. S6).**

In relation with the two additional CRISPR loci detected by CRISPRFinder in SJM180 strain, these were called CRISPR1-like and CRISPR3-like and were found in two different hypothetical proteins (WP_000446591-CRISPR1-like; WP_013356447-CRISPR3-like). Their percentage of identity was not significant for considering that they were the same gene. The CRISPR1-like loci was inserted in the middle of the gene and was located in position 128.894 to 128.614 of this strain's genome, with a length of 314bp, five DRs (with an average length of 25bp), and four spacers (with a length of 34bp). Spacers 1_1 and 3_1 showed similarity with plasmids and viruses, respectively **(Table S2)**. A Blastn search for this gene revealed that this protein is present in 39 strains. Also, it was observed that the sequence from region 5´ to the beginning of CRISPR1-like (approximately 380bp) was highly conserved (91%), while the region corresponding to CRISPR-like was degenerate (63%), with the 3' region (approximately 600bp) being highly conserved (90%) as well **(Fig. S7).** The phylogenetic tree using the 39 CRISPR1-like sequences showed, in this case, a mixture of the strains in relation to their geographical origin. **(Fig. S8)**

The CRISPR3-like region, with a length of 266bp, was also inserted in the middle of the gene (positions 1.201.946 to 1.201.720 of the genome) and showed five DRs (average length of 23bp) and four spacers (ranging between 19 and 31bp), with spacers 1 and 3 showing similarity with plasmids **(Table S2)**. The Blastn analysis revealed this protein to be in 27 strains. This hypothetical protein was highly conserved (96%) from the 5` region to the beginning of the CRISPR-like region (approximately 380bp) while the CRISPR3-like region was degenerate (61%) and the 3` region (approximately 520bp) was conserved (81%) **(Fig. S9).** The phylogenetic tree created with these 27 sequences showed, as in the previous case, a mixture of the strains of different geographical origins **(Fig. S10).**

Cas Domain detection

Cas3 and Cas4 domains were identified in 100% of the analyzed strains, whereas Cas2 domains were found in 32 strains (60.4%), and the Csa3 domain only in 2 strains (4%) **(Table 1, Table S4)**. These domains were found in various locations in the different strains.

Discussion

In the genome of the human pathogenic bacterium Gram negative, H. pylori, the CRISPR-Cas system is not functional and does not exist by forming an operon structure as it is known for other organisms. The lack of this system in some prokaryotes has been related to the increase in the capacity to integrate exogenous DNA in the genome of these bacteria and, resulting in the acquisition of new functions, which can confer an adaptive advantage to these strains, particularly during their transition to pathogenesis [6-33]. But recently, Bangpanwimon et al., 2017 [15] reported the presence of CRISPR-type sequences inserted into the VlpC gene of *H pylori*. In that study, the detection was performed by PCR in partial regions of the genome of Thailand isolates [15]. Each isolated strain showed a CRISPR-like locus with similar DRs sequences. However, results from other strains from different geographical regions, the variability of this locus, or the possibility of the presence of other CRISPR-like loci in the genome of *H. pylori* were not analyzed.

In this work, we show the analysis of 53 strains of *H. pylori* which comprise all the continents. The phylogenetic analyses carried out using the sequences of the CRISPR-like locus found by CRISPRFinder revealed the existence of additional loci to the CRISPR-like locus inserted into the VlpC gene described by Bangpanwimon et al., 2017 [15] **(Fig. S2, S5, S7, and S9)**.

Of the 53 genomes analyzed, 51 of them showed a locus similar to the CRISPR-like locus found in VlpC gene, with DRs and spacer sequences similar to those detected in Bangpanwimon et al., 2017 [15]. In the phylogenetic tree, using the CRISPR-like sequences present in this gene, we observed that the strains that corresponded to an African and European origin formed a differentiated cluster with respect to the Asian and Amerindian strains **(Fig. 3)**. This fact would indicate that the strains furthest from the African origin, such as those of Asian and Amerindian origin, have undergone a process of greater differentiation. Duncan et al., 2013 [34] proposed that the different strains of *H. pylori* were subject to different selective pressures depending on their environmental conditions and according to their phylogeographic origin, and this can lead to the diversification of certain genomic regions, as seems to be the case here [34].

The presence of CRISPR-like loci caused changes in the sequence of the genes where they are inserted into, truncating it or varying its sequence close to the insertion point **(Fig. S2, S5, S7, and S9)**. In addition, the CRISPR loci themselves showed great variability between strains because DRs and spacers were variable in number, even with reverse positions in several genomes **(Fig. S2, S5, S7, and S9)**. These variations indicate recombination phenomena that involve the CRISPR-like locus. In this sense, the CRISPR-like loci could be considered as repetitive sequences involved in intra- and inter- genomic recombination, contributing to the diversity of *H. pylori*. In fact, the variability found between strains, with duplications and deletions within DRs and spacers, could be the result of both types of recombination. In addition, in this work, we showed the presence of a CRISPR-like locus in a region near the breaking point of a large inversion that affects several strains (Shi470 and BM012A), and may therefore be involved in this process (**Fig. S4).** In *Helicobacter* there have been reports about the implications of repeated sequences in this type of rearrangement events [2-35-36]. The implication of CRISPR-like loci in the recombination process could also be supported by the presence of a RecR gene, which is implicated in recombination and repair processes [32], in the same operon as the VlpC gene.

All of these processes would be part of the mechanisms that infer the extreme genome plasticity of *H. pylori* through mutation and recombination intra e inter genomic, exhibiting genetic mosaicism [4].

Currently, it is hypothesized that degenerated CRISPR-Cas systems, or their individual components, as in this case, could derive into diverse roles in a wide range of processes [6]. Thus, if a novel function of a CRISPR system, or one of its components, confers a competitive advantage in the environment in which the organism evolved (that is, it is adaptive) its maintenance and propagation in populations could be a direct result of natural selection.

In fact, it has been shown that orphan CRISPRs loci may be involved in gene regulation. In *Listeria monocytogenes*, orphan CRISPR affected virulence through the FeoAB iron transport system [38]. In this sense, the constant presence of this repeated and mutable structure in these genes of *H. pylori*, and more specifically in the VlpC gene, which is part of the central genome, could be related to the regulation of its expression, as they are located in the promoter region. The integration of the CRISPR-like structure into the VlpC gene would allow the bacteria to be less sensitive to the host defense mechanisms as indicated by Bangpanwimon et al., 2017 [15], and would confer the ability to adapt to different stomach areas, facilitating the capacity to adhere to the gastric epithelium [39]. Similar situations have been described in *Staphylococcus aureus*, in which case the absence of the CRISPR-Cas system conferred the ability to acquire new genes and be more virulent, or as *Enterococus fecalis*, where the modification of their CRISPR-Cas systems made their strains more resistant to antibiotics [6].

Although clustered Cas genes were not detected, and therefore a functional CRISPR-Cas system was also not found, in this work the presence of cas domains in the genome of *H. pylori* was found **(Table S4)**. This presence could signify that the presence of this system is ancestral. This theory could be strengthened by the fact that in other Helicobacter species, CRISPR-Cas systems are present and active [13-14]. The cas domains in the *H. pylori* genome could be performing other functions. In fact, it has been reported, in *H. pylori*, that the VapD protein, associated with a ribonuclease function, is phylogenetically related to Cas2 proteins. Specifically, the HP0315 protein, a member of the VapD family, has a structural similarity to Cas2 and appears to be an evolutionary intermediate between Cas2 and a gene from the Toxin-Antitoxin system [40].

The loss of the functional system is also supported by the fact that in the evolutionary process the number of repetitions present in a CRISPR locus depends on the level of decay of the associated genes [41], as is the case of *H. pylori*, in which the number of DRs observed is low and only cas domains are found, which may be remnants of the original system.

From the analysis carried out, the presence of a CRISPR-like locus within several genes of *H. pylori* was demonstrated. The origin and evolution of these types of sequences is still uncertain. However, for the case of the structure found in the VlpC gene, data is available that has helped with inferring its evolutionary history. In this sense, when comparing the genomes of different *Helicobacter* species, it was found that the VlpC gene was only found in *H. pylori* and *H. cetorum* and with a high degree of similarity. This could indicate that this gene was acquired after the separation of the common ancestor of *H. pylori* and *H. cetorum* from the rest of the species, by duplication from the VacA gene [16]. After this event, the acquisition of the CRISPR-like sequences could have taken place in the VlpC gene. These structures, as pointed out, are in a state of constant flow [42], and therefore they can appear and disappear depending on the selective forces of the environment. During the speciation process of *H. pylori* and *H. cetorum*, the differentiation of CRISPR-like loci occurred between both species. In this sense, it could be said that although the DRs of both species have a high degree of similarity, indicating the common origin, the spacer sequences are variable. It has also been suggested that CRISPR loci can evolve rapidly in some environments, in accordance with the new role played in their antagonistic coevolution [43]. The CRISPR-like loci in *H. pylori* have evolved independently of those of *H. cetorum* (sympatrically), supporting this type of antagonistic coevolution.

In addition, and due to the high degree of change found in these sequences (**Fig. S1**), CRISPR-like loci can be used to determine a strain's origin. Different genomic regions have been used for phylogenetic analyses of *H. pylori* as Multi Locus Sequence Typing (MLST), housekeeping genes and genes of the central genome [44-45-46]. In our case, the phylogeny, using DRs and spacers of CRISPR-like locus within the VlpC gene, groups the strains by geographic origin **(Fig. 3)**, relating the African ones with the European ones, separating them from the Asian and Amerindian ones (of more recent origin). This same situation was observed for the CRISPR-like loci of the Poly-E rich poly genes and for one of the hypothetical proteins (with cell division function), with a grouping by geographical origin **(Fig. S3, and S6).** For this latter protein, the absence of this gene in all strains of the Asian clade was highlighted. Lastly, the sequences of CRISPR1-like and CRISPR3-like loci did not have a geographical grouping (**Fig. S8, and S10)**, showing a process of variation that was independent of the geographical origin.

Conclusions

We detected the presence of different CRISPR-like loci in almost all the analyzed genomes of *Helicobacter pylori* strains with different geographical origins. We characterized their structure as well as their location within the genome. The presence of this type of CRISPR-like causes modifications in the genes where they have been inserted, increasing the variability and in some cases being able to produce genomic rearrangements. On the other hand, its evolution has been associated with geographical regions. Although the function of this type of loci is unknown, several roles have been proposed for this type of structures. For all this, this work highlights the importance of this type of sequences, which seem to have lost their initial function, in the variability and genomic evolution of *H. pylori*.

Competing Interests

The authors declare there are no competing interests.

Availability

The genome sequences of the *H. pylori* and non-*pylori* strains [J99, 2017, 2018, 908, Gambia94, PeCan18, South Africa7, South Africa20, ELS37, HUP-B14, SJM180, P12, B38, G27, UM037, B8, Lithuania75, 26695 (NC_000915), 26695 (NC_018939), Rif1, Rif2, HPAG1, BM012A, BM012S, India7, SNT49, XZ274, OK310, F57, 35A, F16, UM066, UM032, UM299, UM298, F30, OK113, 83, 51, F32, 52, PeCan4, Puno135, Puno120, Sat464, Shi470, Shi169, Shi417, Shi112, Cuz20, v225d, Aklavik86, Aklavik117 and *Helicobacter cetorum* Mit 99-5656] are available in the public database**.**

References

1. Backert, S., & Yamaoka, Y. (Supplementary Fig. S4).2016, Helicobacter pylori research: From bench to bedside. *Helicobacter Pylori Research: From Bench to Bedside*, 1–613.

2. Kang, J., & Blaser, M.J. 2006, Bacterial populations as perfect gases: Genomic integrity and diversification tensions in *Helicobacter pylori*. *Nature Reviews Microbiology*, *4*(11), 826–836.

3. Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., et al (1998). Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(21), 12619–12624.

4. Zawilak-Pawlik A., Zakrzewska-Czerwińska J. 2017, Recent Advances in *Helicobacter pylori* Replication: Possible Implications in Adaptation to a Pathogenic Lifestyle and Perspectives for Drug Design. In: Tegtmeyer N., Backert S. (eds) Molecular Pathogenesis and Signal Transduction by Helicobacter pylori. Current Topics in Microbiology and Immunology, vol 400. Springer, Cham.

5. Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J. 2005, Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. 174–82.

6. Sampson, T.R., & Weiss, D.S. 2013, Alternative Roles for CRISPR/Cas Systems in Bacterial Pathogenesis. *PLoS Pathogens*, *9*(10), e1003621.

7. Grissa, I., Vergnaud, G., & Pourcel, C. 2008, CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *36*(Web Server issue), W145–W148.

8. Zhang, Q., & Ye, Y. 2017, not all predicted CRISPR–Cas systems are equal: isolated *cas* genes and classes of CRISPR like elements. *BMC Bioinformatics*, *18*, 92. http://doi.org/10.1186/s12859-017-1512-4

9. Koonin, E.V., Makarova, K.S., & Zhang, F. 2017, Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology, 37, 67–78.

10. Delaney, N.F., Balenger, S., Bonneaud, C., Marx, C.J., Hill, G.E., Ferguson-Noel, N., et al. 2012, Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum* . *PLoS Genetics*, *8*(2), e1002511.

11. Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., et al. 2016, Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications, 7, 1–8.

12. Burstein, D., Harrington, L. B., Strutt, S. C., Probst, A. J., Anantharaman, K., Thomas, B. C., et al. 2017. New CRISPR-Cas systems from uncultivated microbes. Nature, 542(7640), 237–241.

13. Kersulyte, D., Rossi, M., & Berg, D. E. 2013. Sequence Divergence and Conservation in Genomes of Helicobacter cetorum Strains from a Dolphin and a Whale. PLoS ONE, 8(12), e83177.

14. Tomida, J., Morita, Y., Shibayama, K., Kikuchi, K., Sawa, T., Akaike, T., & Kawamura, Y. (2017). Diversity and microevolution of CRISPR loci in Helicobacter cinaedi. PLoS ONE, 12(10), e0186241.

15. Bangpanwimon, K., Sottisuporn, J., Mittraparp-arthorn, P., Ueaphatthanaphanich, W., Rattanasupar, A., Pourcel, C., & Vuddhakul, V. 2017. CRISPR-like sequences in *Helicobacter pylori* and application in genotyping. *Gut Pathogens*, *9*, 65.

16. Foegeding, N.J., Caston, R.R., McClain, M.S., Ohi, M.D., & Cover, T.L. 2016, An Overview of Helicobacter pylori VacA Toxin Biology. Toxins, 8(6), 173.

17. Albert, T.J., Dailidiene, D., Dailide, G., Norton, J. E., Kalia, A., Richmond, T.A., et al. 2005, Mutation discovery in bacterial genomes: Metronidazole resistance in Helicobacter pylori. Nature Methods, 2(12), 951–953.

18. Bondy-Denomy, J., and Davidson., A.R. 2014, To acquire or resist: the complex biological effects of CRISPR-Cas systems**.** Trends Microbiol., 22 (2014), pp. 218-225.

19. Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. 2010, Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics : TIG, 26(8), 335–340.

20. García-Gutiérrez, E., Almendros, C., Mojica, F.J.M., Guzmán, N.M., & García-Martínez, J. 2015, CRISPR Content Correlates with the Pathogenic Potential of Escherichia coli . PLoS ONE, 10(7), e0131935.

21. Hatoum-Aslan, A., & Marraffini, L.A. 2014, Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. Current Opinion in Microbiology, 0, 82–90.

22. Almendros, C., Guzmán, N.M., García-Martinez, J., & Mojica, F.J.M. 2016, Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. Nature Microbiology, 1(8), 1–8.

23. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., & Sayers, E.W. 2011, GenBank. *Nucleic Acids Research*, *39*(Database issue), D32–D37.

24. Grissa, I., Vergnaud, G., & Pourcel, C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(Web Server issue), W52–W57.

25. Edgar, R.C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

26. Biswas, A., Gagnon, J.N., Bro(Supplementary Fig. S4).uns, S.J.J., Fineran, P.C., & Brown, C.M. 2013, CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biology*, *10*(5), 817–827.

27. Kumar, S., Stecher G., & Tamura., K. 2016, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, *Molecular Biology and Evolution*, Volume 33, Issue 7, 1 July 2016, Pages 1870–1874.

28. Pertea, M., Ayanbule, K., Smedinghoff, M., & Salzberg, S.L. 2009, OperonDB: a comprehensive database of predicted operons in microbial genomes. Nucleic Acids Research, 37(Database issue), D479–D482.

29. Haft, D.H., Selengut, J., Mongodin, E.F., & Nelson, K. E. 2005, A guild of 45 CRISPR-associated (Cas) protein families (Supplementary Fig. S4).and multiple CRISPR/cas subtypes exist in prokaryotic genomes. PloS Computational Biology, 1(6), 0474–0483.

30. Eddy, S. 1998, Profile hidden Markov models. Bioinformatics, 14(9), 755–763.

31. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. 2012, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647–1649.

32. Darling, A.E., Mau, B., & Perna, N.T. 2010, progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, *5*(6), e11147.

33. Sampson, T.R., & Weiss, D.S. 2014, CRISPR-Cas systems: new players in gene regulation and bacterial physiology. Frontiers in Cellular and Infection Microbiology, 4, 37.

34. Duncan, S.S., Valk, P.L., McClain, M.S., Shaffer, C.L., Metcalf, J.A., Bordenstein, S.R., & Cover, T.L. 2013, Comparative Genomic Analysis of East Asian and Non-Asian Helicobacter pylori Strains Identifies Rapidly Evolving Genes. PLoS ONE, 8(1), e55120.

35. Aras, R.A., Fischer, W., Perez-Perez, G.I., Crosatti, M., Ando, T., Haas, R., & Blaser, M.J. 2003, Plasticity of Repetitive DNA Sequences within a Bacterial (Type IV) Secretion System Component. The Journal of Experimental Medicine, 198(9), 1349–1360.

36. Suerbaum, S., & Josenhans, C. 2007, Helicobacter pylori evolution and phenotypic diversification in a changing host. Nature Reviews Microbiology, 5(6), 441–452.

37. Marsin, S., Mathieu, A., Kortulewski, T., Guérois, R., & Radicella, J.P. 2008, Unveiling Novel RecO Distant Orthologues Involved in Homologous Recombination. PLoS Genetics, 4(8), e1000146.

38. Mandin, P., Repoila, F., Vergassola, M., Geissmann, T., & Cossart, P. 2007, Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets. Nucleic Acids Research, 35(3), 962–974.

39. Harvey, V.C., Acio, C.R., Bredehoft, A.K., Zhu, L., Hallinger, D.R., Quinlivan-Repasi, V., et al. 2014, Repetitive Sequence Variations in the Promoter Region of the Adhesin-Encoding Gene sabA of Helicobacter pylori Affect Transcription. Journal of Bacteriology, 196(19), 3421–3429.

40. Kwon, A.R., Kim, J.H., Park, S.J., Lee, K.Y., Min, Y.H., Im, H., et al. 2012, Structural and biochemical characterization of HP0315 from Helicobacter pylori as a VapD protein with an endoribonuclease activity. Nucleic Acids Research, 40(9), 4216–4228.

41. Touchon, M., & Rocha, E.P. C. 2010, the small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. PLoS ONE, 5(6).

42. Marraffini, L.A. 2013, CRISPR-Cas Immunity against Phages: Its Effects on the Evolution and Survival of Bacterial Pathogens. PLoS Pathog 9(12): e1003765.

43. Westra, E.R., Dowling, A.J., Broniewski, J.M., & van Houte, S. 2016, Evolution and Ecology of CRISPR. Annual Review of Ecology, Evolution, and Systematics, 47(1), 307–331.

44. Falush D, Stephens M, Pritchard J.K. 2003, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567-87.

45. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. 2003, Traces of human migrations in *Helicobacter pylori* populations. Science;299:1582-5.

46. Yahara, K., Furuta, Y., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Kobayashi, I, et al. 2013, Chromosome Painting In Silico in a Bacterial Species Reveals Fine Population Structure. *Molecular Biology and Evolution*, *30*(6), 1454–1464.

Figure 1(on next page)

Classification of the repeated consensus sequences obtained from CRISPRFinder.

Phylogenetic tree of there peated consensus sequences obtained from the CRISPR loci confirmed to establish evolutionary relationships and classify these sequences. The MEGA7 software was implemented for this analysis. The evolutionary distance scale is 0.2 Jukes-Cantor model. **(A)** CRISPRlocatedwithintheVlpCgene.**(B)** CRISPR located within genes other than theVlpC gene.

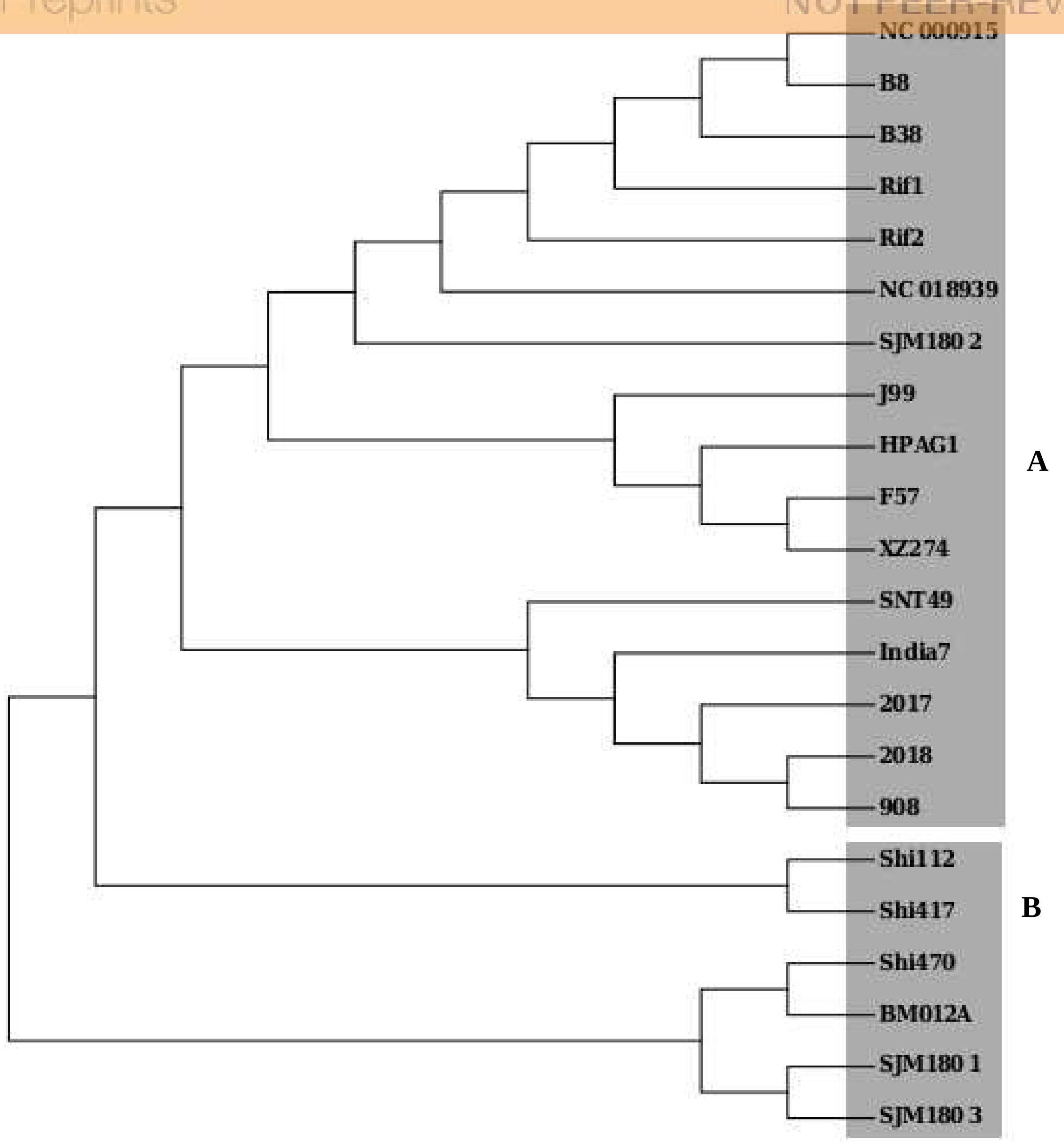
NC 000915
B8
B38
Rif1
Rif2
NC 018939
SJM180 2
J99
HPAG1
F57
XZ274
SNT49
India7
2017
2018
908
A
Shi112
Shi417
Shi470
BM012A
SJM180 1
SJM180 3
B

Figure 2(on next page)

Classification of the spacers sequences obtained from CRISPRFinder.

Phylogenetic tree for the classification of the spacers sequences obtained from the confirmed CRISPR, based on evolutionary relationships implementing the MEGA7 and software. The evolutionary distance scale is 0.1 Jukes-Cantor model**. (A,B,andC)** represent the spacers located within the VlpC gene. **(D)**Represents the spacers located within genes other than VlpC.

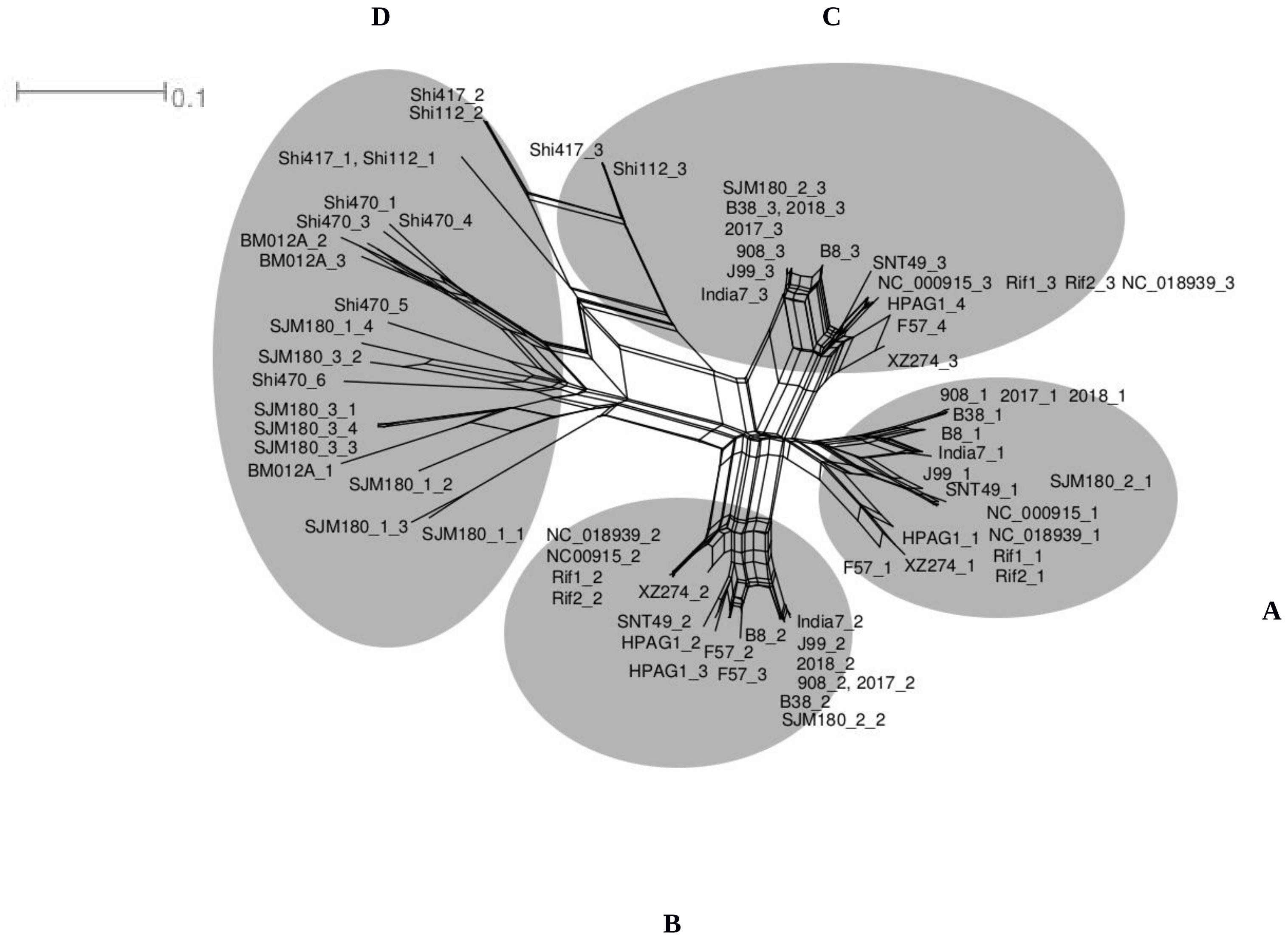
D
C
A
B
0.1
Shi417_2
Shi112_2
Shi417_1, Shi112_1
Shi417_3
Shi112_3
Shi470_1
Shi470_3
Shi470_4
BM012A_2
BM012A_3
Shi470_5
SJM180_1_4
SJM180_3_2
Shi470_6
SJM180_3_1
SJM180_3_4
SJM180_3_3
BM012A_1
SJM180_1_2
SJM180_1_3
SJM180_1_1
SJM180_2_3
B38_3, 2018_3
2017_3
908_3
B8_3
J99_3
India7_3
SNT49_3
NC_000915_3 Rif1_3 Rif2_3 NC_018939_3
HPAG1_4
F57_4
XZ274_3
908_1 2017_1 2018_1
B38_1
B8_1
India7_1
J99_1
SNT49_1
SJM180_2_1
NC_000915_1
HPAG1_1
NC_018939_1
Rif1_1
Rif2_1
F57_1
XZ274_1
NC_018939_2
NC00915_2
Rif1_2
Rif2_2
XZ274_2
SNT49_2
HPAG1_2
F57_2
B8_2
HPAG1_3
F57_3
India7_2
J99_2
2018_2
908_2, 2017_2
B38_2
SJM180_2_2

Figure 3(on next page)

Classification of CRISPR-like in VlpC gene

Phylogenetic tree constructed with the 51 CRISPR-like sequences present and located inside the VlpC gene, which evidences a phylogeographic differentiation of the CRISPR-like loci. Analysis executed with MEGA7 software. The evolutionary distance scales is 0.02 Jukes-Cantor model. **(A)** Group of African and European geographical origin**.(B**) Geographical group of Asian origin and **(C)** Amerind geographic group.

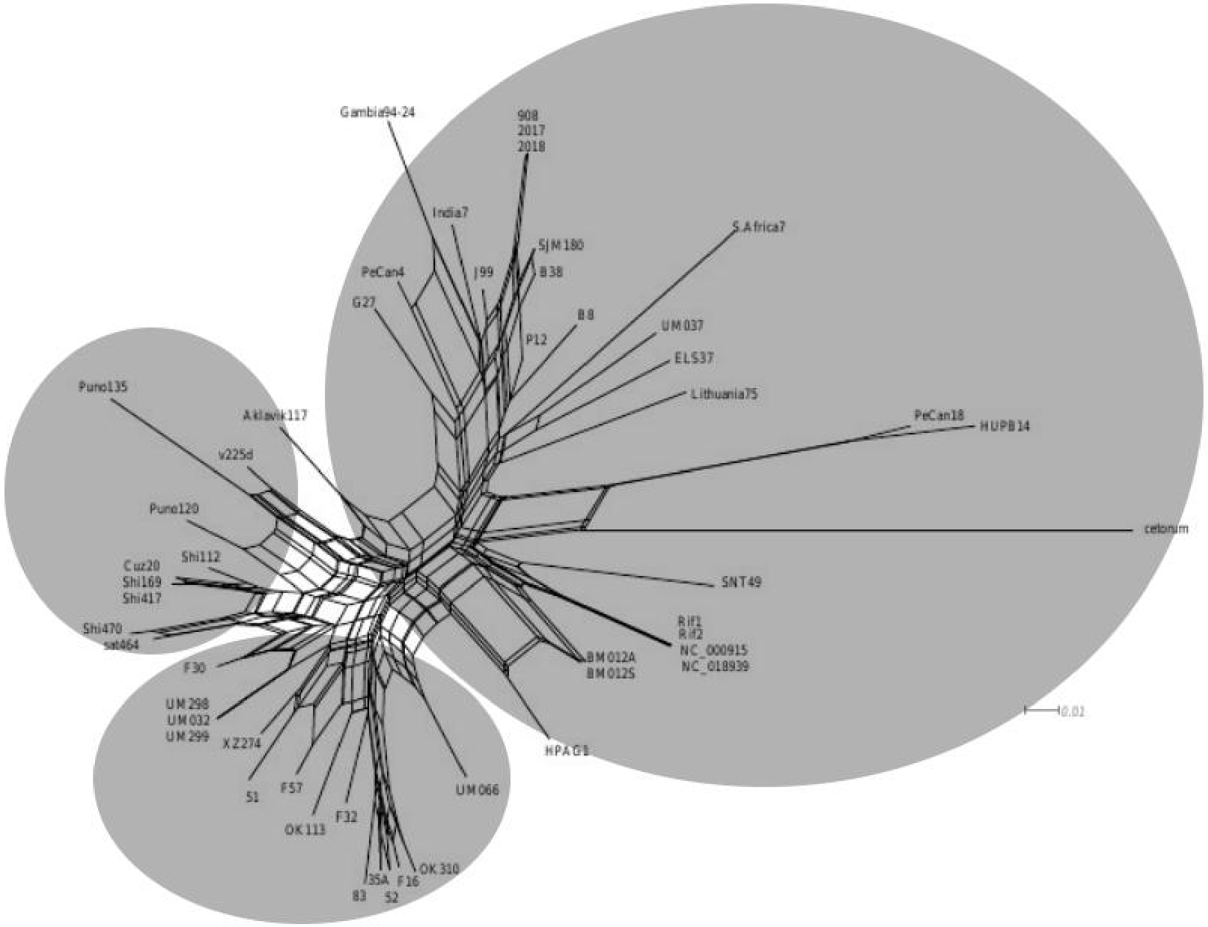
Gambia94-24
908
2017
2018
India7
SJM180
B38
J99
PeCan4
G27
B8
P12
S.Africa7
UM037
ELS37
Lithuania75
PeCan18
HUPB14
cetorum
SNT49
Rif1
Rif2
NC_000915
NC_018939
BM012A
BM012S
HPAG1
Puno135
Aklavik117
v225d
Puno120
Shi112
Cuz20
Shi169
Shi417
Shi470
sat464
F30
UM298
UM032
UM299
XZ274
51
F57
OK113
F32
UM066
35A
F16
OK310
83
52
0.01

Table 1(on next page)

Characteristics of the CRISPR-like loci detected with CRISPRFinder in the 53 strains of *H.pylori*

The different colors represent the presence of cas domains in the analyzed genomes : blue: Cas2, red: Cas3, yellow: Cas4 and green: Csa3. Cas1 domains were no tdetected.

				CRISPRFinder detection	Cas domains detect with Hmmscan (HMMER) E-value 10e-5			
Accesion number	Strain	Origin/isolation	Diagnosis	Gene with CRISPR locus	cas2	cas3	cas4	Csa3
NC_000921	J99	Africa/USA	Duodenal ulcer	VlpC		🟥	🟨	
NC_017374	2017	Africa/France	Duodenal ulcer	VlpC		🟥	🟨	
NC_017381	2018	Africa/France	Duodenal ulcer	VlpC		🟥	🟨	
NC_017357	908	Africa/France	Duodenal ulcer	VlpC		🟥	🟨	
NC_017371	Gambia94/24	Africa/Gambia	unknown			🟥	🟨	
NC_017742	PeCan18	Africa/Peru	gastric cancer		🟦	🟥	🟨	
NC_017361	S. africa7	South Africa	unknown		🟦	🟥	🟨	
NC_022130	S. africa20	South Africa	unknown		🟦	🟥	🟨	
NC_017063	ELS37	America/El Salvador	Gastric cancer			🟥	🟨	
NC_017733	HUP-B14	Europe/Spain	unknown			🟥	🟨	
NC_014560	SJM180	America/Peru	Gastritis	Hypothetical protein -VlpC-Hypothetical protein		🟥	🟨	
NC_011498	P12	Europe/German	Duodenal ulcer		🟦	🟥	🟨	
NC_012973	B38	Europe/France	MALT lymphoma	VlpC		🟥	🟨	
NC_011333	G27	Europe/Italy	unknown			🟥	🟨	
NC_021217	UM037	Asia/Malasya	unknown		🟦	🟥	🟨	
NC_014256	B8	unknown	Gastric ulcer	VlpC	🟦	🟥	🟨	
NC_017362	Lithuania75	Europe/Lithuania	unknown			🟥	🟨	
NC_000915	26695	Europe/UK	Gastritis	VlpC	🟦	🟥	🟨	
NC_018939	26695	unknown	unknown	VlpC	🟦	🟥	🟨	
NC_018937	Rif1	Europe/German	unknown	VlpC	🟦	🟥	🟨	
NC_018938	Rif2	Europe/German	unknown	VlpC	🟦	🟥	🟨	
NC_008086	HPAG1	Europe/Sweden	Atrophic gastritis	VlpC		🟥	🟨	
NC_022886	BM012A	Oceania/Australia	Asymptomatic-reinfection	Poly E-rich protein		🟥	🟨	
NC_022911	BM012S	Oceania/Australia	Asymptomatic-reinfection			🟥	🟨	
NC_017372	India7	Asia/India	Peptic ulcer	VlpC	🟦	🟥	🟨	
NC_017376	SNT49	Asia/India	Asymptomatic	VlpC	🟦	🟥	🟨	
NC_017926	XZ274	Asia/China	Gastric cancer	VlpC	🟦	🟥	🟨	
NC_020509	OK310	Asia/Japan	unknown			🟥	🟨	
NC_017367	F57	Asia/Japan	Duodenal ulcer	VlpC		🟥	🟨	
NC_017360	35A	Asia/Japan	unknown			🟥	🟨	
NC_017368	F16	Asia/Japan	Gastritis			🟥	🟨	

NC_021218	UM066	Asia/Malasya	unknown			🟥	🟨	
NC_021215	UM032	Asia/Malasya	peptic ulcer			🟥	🟨	
NC_021216	UM299	Asia/Malasya	unknown			🟥	🟨	
NC_021882	UM298	Asia/Malasya	unknown			🟥	🟨	
NC_017365	F30	Asia/Japan	Duodenal ulcer			🟥	🟨	
NC_020508	OK113	Asia/Japan	unknown		🟦	🟥	🟨	
NC_017375	83	unknown	unknown		🟦	🟥	🟨	
NC_017382	51	Asia/Korea	Duodenal ulcer			🟥	🟨	
NC_017366	F32	Asia/Japan	Gastric cancer		🟦	🟥	🟨	
NC_017354	52	Asia/Korea	unknown		🟦	🟥	🟨	
NC_014555	PeCan4	America/Peru	gastric cancer		🟦	🟥	🟨	
NC_017379	Puno135	America/Peru	Gastritis		🟦	🟥	🟨	
NC_017378	Puno120	America/Peru	Gastritis		🟦	🟥	🟨	
NC_017359	Sat464	America/Peru	unknown		🟦	🟥	🟨	
NC_010698	Shi470	America/Peru	Gastritis	Poly E-rich protein	🟦	🟥	🟨	
NC_017740	Shi169	America/Peru	unknown		🟦	🟥	🟨	
NC_017739	Shi417	America/Peru	unknown	Hypothetical protein	🟦	🟥	🟨	
NC_017741	Shi112	America/Peru	unknown	Hypothetical protein	🟦	🟥	🟨	🟩
NC_017358	Cuz20	America/Peru	unknown		🟦	🟥	🟨	🟩
NC_017355	v225d	America/Venezuela	Gastritis		🟦	🟥	🟨	
NC_019563	Aklavik86	America/Canada	Gastritis		🟦	🟥	🟨	
NC_019560	Aklavik117	America/Canada	Gastritis		🟦	🟥	🟨	