Comparative Genomic Analysis of Statistically Significant Genomic Islands of Helicobacter pylori strains for better understanding the disease prognosis

Joyeeta Chakraborty¹, Raghunath Chatterjee¹#

¹Human Genetics Unit, Indian Statistical Institute, 203 B T Road, Kolkata, India 700 108

Address correspondence to:
Dr. Raghunath Chatterjee (email: rchatterjee@isical.ac.in)
Human Genetics Unit,
Indian Statistical Institute,
203 B T Road,
Kolkata, India 700 108

Running title: Genomic Islands in Helicobacter pylori
Abstract

Bacterial virulence factors are often located in their genomic islands (GIs). *Helicobacter pylori*, a highly diverse organism is reported to be associated with several gastrointestinal diseases like, gastritis, gastric cancer, peptic ulcer, duodenal ulcer etc. A novel similarity score-based comparative analysis with GIs of fifty *H. pylori* strains revealed clear idea of the various factors which promote disease progression. Two putative pathogenic GIs in some of the *H. pylori* strains were identified. One GI, having a putative labile enterotoxin and other dynamin-like proteins (DLPs), is predicted to increase the release of toxin by membrane vesicular formation. Another island contains a virulence-associated protein D (vapD) which is a component of a type-II toxin-antitoxin system (TAs), leads to enhance the severity of the *H. pylori* infection. Besides the well-known virulence factors like CagA, and VacA, several GIs have been identified which showed to have direct or indirect impact on *H. pylori* clinical outcomes. One such GI, containing lipopolysaccharide (LPS) biosynthesis genes was revealed to be directly connected with disease development by inhibiting the immune response. Another collagenase-containing GI worsens ulcers by slowing down the healing process. GI consisted of *fliD* operon was found to be connected to flagellar assembly and biofilm production. By residing in biofilms, bacteria can avoid antibiotic therapy, resulting in chronic infection. Along with well-studied CagA and VacA virulent genes, it is equally important to study these identified virulence factors for better understanding *H. pylori* induced disease prognosis.

Keywords: Genomic Island, Horizontal Gene Transfer, Comparative Genomics, Similarity Score, Toxin-antitoxin system, Pathogenicity Island.
Introduction

*Helicobacter pylori*, formerly known as *Campylobacter pylori*, is a Gram-negative, microaerophilic bacteria that is part of the gastric microbiota in over 50% of population worldwide. The International Agency for Research on Cancer (IARC) classified it as a Class I carcinogen in 1994, nearly 20 years after its discovery, stating that people having *H. pylori* infection have six times higher risk of developing gastric cancer (GC) than those who do not [1]. Colonization of *H. pylori* in stomach is commonly acquired in early age, mostly by person-to-person transmission through fluidic or aerosolic transfer mediated by the faeco-oral route [2, 3]. *H. pylori* infection leads a variety of upper gastrointestinal disorders, such as chronic gastritis, peptic ulcer disease (PUD), gastric mucosa-associated lymphoid tissue (MALT) lymphoma, and GC [4-6]. Colonization persists throughout life unless treated with antibiotics. Unfortunately, the spike in antibiotic resistance is clearly affecting the response to the therapy and there is currently no preventive vaccine measure. Different strains of the *H. pylori* were found to have different level of virulence, while some of the strains were found to be non-virulent. So, to understand the disease prognosis in a better way and to develop new therapeutics to fight against this bacterium, it is important to study *H. pylori* in strain level. Strain specific virulence factors are generally located in the Genomic Island (GI) region of particular pathogenic strains. Large genomic regions consisting of multiple horizontally transferred genes (HTGs) are collectively termed as genomic islands (GIs) [7]. These are crucial for bacterial evolution, adaptability, and diversification as they contain genes that code for a wide range of proteins [8]. In 1990 Hacker et al. identified virulent genes in uropathogenic strains of *Escherichia coli*. This identified virulence factor was not present in beneficial strains of *E. coli*. This cluster of genes was referred to as pathogenicity islands (PAI). It has been proposed that possible vaccine candidates are located within PAIs [9]. Several other reports suggested apart from carrying virulent factors, GIs can have clusters of genes related to special biological functions such as, genes required for adaptation in a special environment, genes encoding proteins linked to metabolic processes, antimicrobial resistance genes. Based on the function of the genes present on GIs, they can be differentiated as metabolic islands (MIs), resistance islands (REIs) etc [10, 11].

Despite the fact that *H. pylori* infection is a leading cause of GC, the majority of people who carry the bacteria in their stomachs never get diagnosed with GC. It was also found that some of the normal individuals showed to have higher abundance of *H. pylori*. They neither showed any symptoms nor had any discomfort. *H. pylori* showed to be the part of normal
microflora for these samples. So the role of *H. pylori* in GC prognosis is still unclear. *H. pylori* is a highly diverse organism. Studying *H. pylori* at strain level may shed light to understand the disease prognosis. In the present study the GIs were thoroughly studied in fifty *H. pylori* strains in order to find strain-specific novel virulent factors and other Islands. The comparative study with the GIs among fifty *H. pylori* strains identified GIs with virulence-associated genes in addition to the known GIs. Further characterization of the newly detected islands identified the possible function of these GIs. Several other factors were also identified which were found to be associated either directly or indirectly with the *H. pylori* induced disease prognosis.

**Methods**

**Comparative analysis of GIs in Helicobacter pylori genomes**

Fifty completely sequenced strains of *H. pylori* were downloaded from ftp://ftp.ncbi.nlm.nih.gov/. All of these strains have common hosts i.e. human. Geographical distribution and their disease association were listed in **Supplementary Table 1 and 2**.

*Design-Island-II* (https://www.isical.ac.in/~rchatterjee/Design-Island.htm), a de novo algorithm for GI predication was run on 50 *H. pylori* strains to predict the statistically significant GIs. Islands that were ≥10kb in length were annotated as GIs. All HTGs within these GIs were identified from the gene annotation data using an in-house perl script. For comparison of the GIs among different strains, we used cd-hit suite (http://weizhong-lab.ucsd.edu/cdhit-web-server/) to assign unique Cluster numbers (CLS) to the HTGs. Sequences with 90% identity were given identical cluster (CLS) numbers.

**Similarity Score (Sm) based visualization tool**

To identify the similarity of a particular GI among different *H pylori* strains, we introduced a similarity score (Sm):

\[ Sm = \frac{A \cap B}{A \cup B} \]

where, A and B correspond to number of genes within the given GIs as identified by *Design-Island-II*. Similarity score (Sm) calculates the number of commons genes present in the two GIs, divided by the number of total genes present in the two GIs. We calculated Sm score for all the identified GIs in fifty *H. pylori* strains. Sm value-based heat map were generated for all GIs to get an idea about the similarities or dissimilarities among the fifty *H. pylori* strains.
Hierarchical clustering on the Sm score and heat maps were generated for all GIs using R packages (gplots and heatmap.2).

Core and Accessory Island
A cut-off value of $Sm \geq 50\%$ was used to annotate a GI as shared between two strains. GIs present in all 50 strains, were classified as core GIs. These were named as Core_1, Core_2, etc. Accessory GIs were defined as GIs found in 2 to 49 $H. pylori$ strains and termed as Acc_1, Acc_2, Acc_3 and so on. The presence and absence of GIs across all the strains were denoted as 1 and 0. Kendall's Tau correlation coefficient was calculated among all 50 $H. pylori$ strains considering the presence or absence of accessory islands.

Annotation of hypothetical proteins
A large number of genes within many accessory islands were annotated as hypothetical proteins. Functional characterization of these unknown proteins was performed using a conserved domain search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and psi blast (NCBI) against PDB database with default parameters until convergence. The query sequences were searched against CDD v3.19 - 58235 PSSMs database, using 0.010 as expect value. To understand the conservation of the homolog in the organisms and their evolutionary relationships, multiple sequence alignment and phylogenetic analysis were carried out using Molecular Evolutionary Genetic Analysis (MEGA) [12].

Protein-protein interaction and Homology modeling
The probable biological role of the uncharacterized proteins is predicted by studying the protein–protein interaction using STRINGv10 [13]. Protein three-dimensional structures were generated using an automatic fold recognition server Phyre2 [14]. The PDB IDs of the templates used for homology modeling of putative enterotoxin, vapD and DUF3240 family protein were 4AUR (Confidence: 99.9%; Coverage: 95%), 3UI3 (Confidence: 100.0%; Coverage:96%) and 3CE8 (Confidence: 93.5%; Coverage: 75%) respectively. Clashes, rotamers and Ramachandran analysis were done by Molprobity [15]. Fpocket2 and CSA programs were used to detect pocket and catalytic site of the hypothetical protein respectively [16, 17].Visualization and modification of the 3D model were carried out in PyMOL [18].

Docking
Protein-peptide interactions are essential in various biological processes involving signaling, cellular localization, etc. Twenty-six probable complexes were generated using a protein-protein docking server ClusPro [19]. The complex that had the highest binding energy and lowest free energy was considered. In a protein complex some of the residues take part in the
interaction. BIPSPI was used to predict partner-specific binding sites in the protein-protein complex [20]

Analysis of GIs with other GI prediction tools
To validate the identified GIs with other GI prediction tools, sequence composition-based approaches like, PredictBias [21], Zisland Explorer [22], MSGIP [23]; comparative genomics approach like, IslandPick [24]; and hybrid approach like IslandViewer 4 [25] were selected. These methods were applied on fifty \textit{H. pylori} strains to identify the GIs and their locations in the genome. The length of the predicted GIs were compared with that of the GIs predicted by \textit{Design-Island-II}. If at least 50\% of the length of a GI, predicted by any other tool was overlapped with our predicted GIs, it was determined as detected by both methods.

Results
Identification of GIs across fifty \textit{H. pylori} strains
\textit{Design-Island-II}, a modified version of \textit{Design-Island} [26] was used to detect statistically significant GIs in all fifty \textit{H. pylori} strains. A total of 41 GIs with $\geq 10\text{Kb}$ of length were included for the comparative analysis of fifty strains. The \textit{H. pylori} 35A strain had the highest number of GIs (36 GIs), followed by two Fukui strains (\textit{H. pylori} F16 and \textit{H. pylori} F30) and \textit{H. pylori} J99, both of which had 35 GIs. The Malaysian strains \textit{H. pylori} UM037 and \textit{H. pylori} UM066 had the lowest number of predicted GIs (25 GIs) (Figure 1A). Heat map using the Kendall's Tau correlation coefficient with the GIs among all fifty strains showed clustering of strains according to their place of isolation. The higher degree of correlation among geographically related strains suggested a region-specific inter-strain HGT among strains. Like, Australian strains \textit{H. pylori} BM012A and \textit{H. pylori} BM012S shared almost identical GIs distribution between them, but these were not so similar to other strains isolated from other geographical locations. South East Asian strains, specifically the Malaysian strains (\textit{H. pylori} UM032, \textit{H. pylori} UM037 and \textit{H. pylori} UM066) were quite different from each other and other geographical strains. On the contrary, the strains from Japan, Korea and France showed $\geq 0.80$ correlation among themselves. South African strains \textit{H. pylori} 908, \textit{H. pylori} 2017 and \textit{H. pylori} 2018 were almost identical in terms of shared GIs (Figure 1B).

Comparative analysis of GIs in \textit{Helicobacter pylori}
Forty-one GIs identified in all 50 strains were subjected to comparative analysis among all \textit{H. pylori} strains. Among these, seven GIs were present in all 50 strains and were termed as core GI. Three core GIs (Core_1-Core_3) were found to have $Sm \geq 80\%$. The rest four core GIs
Thirty four accessory GIs (namely, Acc_1 - Acc_34) were present in 2 to 49 *H. pylori* strains with \( Sm \leq 80\% \) (Figure 1C). The plausible functions of all the forty one GIs were predicted (Supplementary Table 3). Two of the seven core GIs were mainly enriched with ribosomal proteins. In terms of composition, many highly expressed genes, like ribosomal proteins, deviate from the genomic background [27, 28]. Thus, they were identified as GIs despite the fact that they were part of the host genome. Various studies suggested eliminating ribosomal proteins containing GIs, as these are false positive in terms of identification [29, 30].

**Characterization of Identified Core Genomic Islands**

**Antibiotic Resistance Island**

Core_3 was found to be present in all *H. pylori* strains with an average \( S_m = 83.1\% \). The heatmap gave a lucid pictorial illustration of the similarity of Core_3 GI among strains (Supplementary Figure 1A). The length of Core_3 was \(~17\text{Kb}\) and 14 HTGs were present on it. This gene cluster was the entire nuo operon (NADH: ubiquinone oxidoreductase). This operon contains essential genes with unique genomic properties, such as different oligonucleotide frequencies and slightly greater GC\% than tissue-specific genes differentiate this region from rest of the genome [31, 32]. Thus, this nuo operon was identified as GI. Benzimidazole (BI) derivatives resistant *H. pylori* strains have three mutations in NADH-quinone oxidoreductase subunit D (NuoD) at G398S, F404S, and V407M and one in NADH-quinone oxidoreductase subunit B (NuoB) at T27A. Despite the fact that this GI contains essential genes, it can be classified as an antibiotic-resistant island. The synteny of fourteen genes in this GI revealed a framework in which homologous gene conservation and gene order were observed (Supplementary Figure 1B).

**Vacuolating toxin A (vacA) PAI**

Core_4, a \(~17\text{Kb}\) long GI was detected to contain vacA gene along with another toxin flippase-type exporter (*mviN*), iron (III) dicitrate ABC transporter genes *fecD* and *fecE*, short-chain oxidoreductase (*vdlC*) in almost all. VacA was deleted in five *H. pylori* strains like, *H. pylori* Shi169, *H. pylori* UM037, *H. pylori* Rif1, *H. pylori* Rif2, *H. pylori* SouthAfrica7. In *H. pylori* 52 and *H. pylori* BM012S vacA was annotated as non-coding. Synteny of this GI revealed that it differed between strains depending on gene content (Supplementary Figure 2A) with average \( Sm \) score of \( 77\% \). The hierarchical clustering based on the similarity score grouped the samples in two major clusters (Supplementary Figure 2B). This GI is associated with virulence as well as iron transport. *H. pylori* Rif1, *H. pylori* Rif2 strains
lacked vacA gene, though in the other vacA lacking strains, vacA paralogs were present on another accessory GI, Acc_23.

**Niche Adaptation Island (Urease)**

The urease gene cluster (Core_5), which was adopted horizontally, became one of the most important genetic elements for the survival of helicobacter strains. The GI was ~17.5Kb in length and consisted of 16 HTGs. Genes encoding *H. pylori* urease are located as a single 6.13kb gene cluster that consists of seven contiguous genes ureA, ureB, ureI, ureE, ureF, ureG and ureH, are necessary for the synthesis of an active enzyme [33, 34]. The average Sm was found to be 58%. The heatmap with the Sm of Urease Island showed three distinct clusters (Figure 2A). Three major clusters were observed in this GI. The insertion or deletion of genes around the conserved seven genes resulted in three different clusters. There were insertion of hypothetical genes and non-coding sequence in the downstream region of the conserved urease gene cluster. Malaysian strains (*H. pylori* UM032, *H. pylori* UM037, *H. pylori* UM066), Venezuela strain *H. pylori* V225, Fukui strain *H. pylori* F57 showed distinct gene arrangements (Figure 2B). The effective colonization by the bacteria in the host is the first stage towards conferring pathogenicity.

**PAI containing Collagenase**

Core_6, a 13Kb long GI, was identified with genes that promote the virulence of pylori infection. This GI was present in all strains, whether intact, fragmented, or missing in part. However, a U32 family peptidase was present in all strains. Other important HTGs of this GI were a DNA-binding response regulator (racR), peptide chain release factor RF2 (prfB), flagellar biosynthetic protein FliR (fliR), fructose-bisphosphate aldolase (fbaB), etc. The average Sm, calculated on the basis of gene content was found to be 64.5%. Thirty-six strains showed >80% similarity and formed the main cluster in the hierarchical clustering (Supplementary Figure 3A). U32 family peptidase present in all the strains is actually a collagenase. Gastric epithelium extracellular matrix is mostly composed of collagen type I and III. Type I collagen synthesis is increased in the areas surrounding gastric ulcers. Thus, ulcer healing process is promoted [35]. However, the release of collagenase from *H. pylori* inhibits the healing process, thus increase the chronicity of the ulceration. Synteny of this GI showed all the fifty strains have this virulence factor, but the difference in the gene content of the adjacent region led to formation of distinct clusters (Supplementary Figure 3B).

**Niche adaptation (Biofilm formation)**
Core_7 was found to contain fliD operon along with some other HTGs, like, DNA-damage induced multidrug efflux protein (dinF), N-carbamoylputrescine amidase (cpa), 2', 3'-cyclic-nucleotide 2'-phosphodiesterase (ymdA) etc. This GI was ~12Kb long and was present in all strains. The fliD operon not only promotes H. pylori colonization in the human stomach by forming flagella, but it also facilitates biofilm formation [36]. The bacteria can evade antibiotic therapy by living in a biofilms, leading to chronic infection [37]. This operon was intact across all the fifty H. pylori strains; with 75% of average Sm score (Supplementary Figure 4A). The difference in adjacent gene cluster led to the formation of two major groups (Supplementary Figure 4B). Core_7 is involved in H. pylori colonization and biofilm formation, which contributes to the disease's chronicity. As fliD is present in all strains and crucial for H. pylori survival, it can be a potential target for new drug or vaccine development.

**Accessory Genomic Islands associated with virulence**

**Lipopolysaccharide (LPS) Biosynthesis Island**

Acc_17 GI was ~13Kb in length and contained ten HTGs present on it, like, lipid-A-disaccharide synthase (lpxB), hof-family outer membrane protein (hofC, hofD), CDP-diacylglycerol diphosphatase (cdh) etc. This GI was found in 41 strains, with an average Sm of 55.3%. The hierarchical clustering in the heatmap divided the strains into two major clusters (Figure 3A). Synteny of this GI also represented that the gene cluster was not intact in all strains, and it was fragmented or some genes were missing (Figure 3B). This GI was found to be associated with lipopolysaccharide (LPS) synthesis. Lipid A-core and the O antigen polysaccharide make up LPS which serves as a key component on the surface of H. pylori. Lewis antigens are present in the O polysaccharides which imitate the glycan structures produced by human cells. Lewis antigens interact with human dendritic cells, generating an immune response that contributes to H. pylori pathogenicity [38]. As LPS is involved in producing H. pylori induced virulence, this GI may indirectly influence the pathogenesis of the strain.

**Cytotoxin associated gene A (CagA) PAI**

Cag is an exhaustively studied pathogenicity island in H. pylori. Comparative analysis of the cagPAI (Acc_25) identified ~80% of the strains have 70-100% of resemblance among them (Supplementary Figure 5A). It was ~40kb in length and consisted of thirty two genes. Three strains (H. pylori 2017, 2018 and 908) showed ~90% similarity among themselves however differed from the other strains. These three strains, reportedly associated with Duodenal Ulcer (DU), belong to the same geographical location i.e. Africa. While five Cag negative strains
H. pylori Aklavik 86, H. pylori Aklavik 117, H. pylori B38, H. pylori SA7 and H. pylori SA20 were clustered together. Hierarchical clustering based on the similarity score, Sm, generated from gene content showed three distinct clusters. Forty-five strains Cag positive strains were either associated with a specific disease or isolated from patients. Four Fukui strains (H. pylori F16, H. pylori F30, H. pylori F32 and H. pylori F57), two Australian (H. pylori BM012A and H. pylori BM012S), two Germany (H. pylori Rif1 and H. pylori Rif2) and two Korean (H. pylori 51 and H. pylori 52) strains showed to have identical gene content among themselves, suggesting an intra-species horizontal transfer of CagPAI within a particular geographical location. Synteny of this island showed three representatives from three distinct clusters observed in the heatmap (Supplementary Figure 5B). Gene composition analysis of the cagPAI suggested additive or reductive evolution in some of the H. pylori strains.

Lysine Biosynthesis Island

Another biosynthetic island (Acc_14) was identified in thirty-three H. pylori strains. The GI was ~15Kb long and consisted of eleven HTGs. Diaminopimelate decarboxylase (lysA), chorismate mutase, amidase (amiE), para-aminobenzoate synthase glutamine amidotransferase (pabB) were some of the genes which are linked to the lysine biosynthetic process. The similarity among the strains was $S_m = 69\%$. Hierarchical clustering identified three distinct clusters. The gene content of Acc_14 in H. pylori 51, H. pylori 83 and H. pylori Lithuania 75 were quite different from others (Supplementary Figure 6A). The functional relevance of this particular GI came up with an interesting finding. Previous reports suggested consumption of a lot of salt in the diet has been linked to an elevated risk of stomach cancer [39-42]. CagA, amiE, lysA, IgE2, and other proteins were found to have increased expression in bacteria cultivated in high-salt conditions [43]. So, this led to the conclusion that the strains containing amiE, lysA enhance the risk of GC in people who take a high-salt diet. The following strains H. pylori 35A (HMPREF4655_20533), 51 (KHP_0287), H. pylori 908 (hp908_0304), H. pylori 2017 (hp2017_0297), H. pylori B8 (HPB8_1272), H. pylori B38 (HELPY_0296), H. pylori F30 (HPF30_1004), H. pylori F32 (HPF32_0300), H. pylori F57 (HPF57_0344), H. pylori J99 (jhp0275), H. pylori OK113 (HPOK113_0299), H. pylori OK310 (HPOK310_0296), H. pylori P12 (HPP12_0289), H. pylori SA20 (HPSA20_0322), and H. pylori V225D (HPV225_0308) have lysA and amiE, which thus increase the chance of GC prognosis. However, lysA and amiE were absent in H. pylori Cuz20 and H. pylori Aklavik86 respectively (Supplementary Figure 6B).

Metabolic Island
Acc_20 was discovered in forty-eight H. pylori strains, spanning 12.6Kb in length and including ten HTGs. Two genes found on this GI were primarily linked with the citric acid cycle: 3-oxoacid CoA-transferase A and 3-oxoacid CoA-transferase B (Sco A and B; known as SCOT complex). Other proteins found on it include acetone carboxylase subunit gamma (acxC), hydantoin utilisation protein A (hyuA), and acetyl-CoA C-acetyltransferase (thiA), short-chain fatty acids transporter (atoE), polysaccharide biosynthesis protein (wlaX) etc. On the basis of the gene content the average Sm was calculated to be 53.8%. All of the studied strains had scoA, scoB and acxC present on Acc_20. Variations in the neighboring gene content led to formation of four separate clusters among the strains (Supplementary Figure 7A-B).

Putative labile enterotoxin containing PAI

HTGs of several accessory GIs were annotated as hypothetical proteins. While characterizing the function of these GIs, we identified a GI (Acc_31) consisted of eleven hypothetical proteins with a labile enterotoxin output A. This labile enterotoxin was found only in five pathogenic strains, namely H. Pylori 908, H. Pylori 2017 and H. Pylori 2018, isolated from DU patients. H. pylori Puno135 and H. pylori F32 were isolated from gastritis and GC patients respectively. Synteny study of these strains differed from other H. pylori strains (Figure 4A & 4B), suggesting its role in disease pathogenesis. These 5 strains were grouped together as having a similarity of ~95% however; gene content of this cluster was quite dissimilar with other strains (Figure 4C). Apart from the putative labile enterotoxin output A, this island consisted of three DLP helical superfamily proteins, GTPase, apolipoL family protein, cyclic beta-1,2-glucan ABC transporter, RsgAGTPase superfamily and ATP-binding protein. Phylogenetic analysis showed that the closest homolog of this toxin was present in another epsilonbacteria Campylobacterjejuni. Other close homologs were found in genus like Proteus, Pseudomonas, Haemophilus and Vibrio (Figure 5A). Protein-protein interaction study showed that the putative labile enterotoxin output A was co-expressed and present in the neighborhood of tRNA(Ile)-lysidine synthase (tilS). TilS co-occurred and co-expressed with D-alanine-D-alanine ligase (ddl) and tRNA-dihydouridine synthase (dus) (Figure 5B). Ddl was associated with peptidoglycan biosynthesis and vancomycin resistance whereas dus was associated with tRNA processing. The Protein 3D model of the putative toxin was generated by homology modeling. Ramachandran plot, rotamer study, clash analysis of the built 3D structure gave an idea that it modeled ideally (Supplementary Figure 8A-C). The binding pocket, where it binds with its ligands, was made up of 26 amino acids. Fifty percent residues of this pocket were found to be conserved among various closely related bacteria.
The homology models of the putative toxin and its binding pocket is presented in Figure 5D & 5E.

**PAI containing Toxin-Antitoxin system (TAs)**

*Design-Island-II* identified a PAI (Acc_33), consisted of a Toxin-Antitoxin system (TAs) in 29 of the studied strains. This GI contains a virulence-associated protein D (*vapD*), Cobalt-zinc-cadmium resistance protein, cation system protein, and six proteins with unknown function. These hypothetical proteins were found to be either dynamin or DUF3240 family protein. VapD is a toxin with purine-specific endoribonuclease activity whereas DUF3240 family protein, with unknown function, binds to its own regulatory region. The toxin and antitoxin were found to be present on the GI of all strains; however, the neighboring genes within the GIs were missing in 26 strains except *H. pylori* F16, *H. pylori* F30 and *H. pylori* F32 (Figure 6A). In type II TAs, the toxic protein is inhibited post-translationally by binding of a less stable antitoxin protein. VapD and its antitoxin were modeled based on their available homologs. For proper functioning, the toxin-antitoxin components must interact physically through electrostatic forces, hydrogen bonding or hydrophobic effect. To check toxin-antitoxin structural interaction docking is performed. The residues responsible for protein-protein interaction in the docked complex were identified in both vapD and its partner (Figure 6 B-C). Among twenty-six protein complexes; the model having the best interaction energy is presented in Figure 6D-E.

**Prediction of GIs with other GI prediction tools**

*Design-Island-II* detected a total of thirty-four accessory GIs in fifty strains of *H. pylori*. Out of which twenty-three GIs were also identified by at least one of the other five GI prediction tools (*Supplementary Table 4*). One GI (Acc_25) was predicted by all five methods, while thirteen GIs were predicted by three methods (*PredictBias* [21], Zisland Explorer [22], and MSGIP [23]), and five GIs were predicted by two, and four GIs were predicted by two GI prediction tools (*Supplementary Table 4*). Two newly identified putative pathogenic GIs, Acc_31 and Acc_33 were detected by four and three GI prediction tools respectively. Acc_31 (labile enterotoxin containing PAI) was detected in all five strains by these four methods, while Acc_33 (Toxin-Antitoxin system PAI) were detected in twenty-nine strains by the *Design-Island-II* and MSGIP. PredictBias, on the other hand, identified Acc_31 in twenty-one strains. A synteny analysis revealed that the Toxin-Antitoxin system were present in all strains, but PredictBias was unable to identify this GI, probably because to the lack of surrounding gene content (*Supplementary Figure 9*).
Eleven accessory GIs were identified only by Design-Island-II. Among these, four GIs (Acc_9, Acc_15, Acc_26, and Acc_34) contained ribosomal proteins along with some HTGs of miscellaneous functions. Many highly expressed genes, such as ribosomal proteins, diverge from the genetic background in terms of composition. As a result, despite the fact that they were essential genes, they were identified as GIs. Several studies have proposed that GI-containing ribosomal proteins be eliminated since they are false positives in terms of identification [29, 30].

Discussion

Comparative analysis among fifty strains of *H. pylori* taking led to identify some new statistically significant GIs in addition to the known GIs. Characterization of the detected GIs gave an idea of their impact in the *H. pylori* colonization, infection and disease prognosis. The nuo operon encoding NADH: ubiquinone oxidoreductase or complex I was found as a GI in all fifty strains. Complex I, present in inner mitochondrial membrane, have multi subunits and responsible for catalyzing the first step of cellular respiration. Essential genes serve as lucrative targets of drug development. The standard *H. pylori* eradication therapy is done using PPIs along with antibiotics like, amoxicillin, metronidazole, clarithromycin. Benzimidazoles (BIs) were found to have potent anti *H. pylori* effect in metronidazole or clarithromycin resistant strains. However, three missense mutations at amino acid position G398S, F404S, V407M in NuoD and one point mutation at T27A in NuoB led to the evolution of BIs resistant strains [44].

The effective colonisation by the bacteria in the host is the first stage towards conferring pathogenicity. The urease gene cluster is necessary not only for establishing colonization but also for maintaining infection of *H. pylori*. This multimeric enzyme, made up of 12 UreA and UreB heterodimers, catalyses the hydrolysis of urea to create CO₂ and NH₃. It serves as a buffer against the acidity of the surrounding environment. [45]. CO₂ and NH₃, diffuse to the periplasmic region where alpha-carbonic anhydrase converts CO₂ in the bicarbonate (HCO₃⁻) and H⁺. The HCO₃⁻ maintains the periplasmic pH close to 6.1, assisting *H. pylori* to avoid acidic pH [46]. The H⁺ ions, responsible for imparting acidity is also taken up by NH₃ in the periplasmic region, another way of acid acclimation [47]. As a result, *H. pylori* combat the deadly effect of the acidic pH. CagPAI is a well-studied virulence factor in *H. pylori* [48]. The presence of the cagA gene, though not a virulence determinant, makes a strain more virulent by eliciting the production of interleukin (IL)-8 by gastric epithelial cells [49].
CagPAI encodes \textit{cagA} an onco-protein which gets transferred to the epithelial cells through T4SS. \textit{VacA} is known to form vacuole, facilitates \textit{H. pylori} colonize and a pore-forming toxin. The recruitment of inflammatory cells by \textit{VacA} promotes ulceration. \textit{VacA} mutant strains were found to produce stomach ulcers less likely than wild-type strains in gerbils [50]. CagA, and VacA gene containing GIs have direct impact on GC prognosis, after successful colonisation of \textit{H. pylori} in the human stomach mucosa with the help of urease gene cluster containing GI. Thus these three GIs play significant role in the development of \textit{H. pylori}-infection mediated illnesses.

After the successful colonization \textit{H. pylori} damages the insulating mucosal barrier of the stomach and duodenum, allowing acid to penetrate the delicate lining beneath. The acid and germs both damage the lining, resulting in a sore or ulcer. The rapid regeneration of the epithelium layer is not the only mechanism for healing injury; fibrillar collagens of types I and III have been reported to be over expressed [35]. These simultaneous healing processes are found in gastric or duodenal ulcer. Core_6 was found to contain a U32 family protein, which is a collagenase in nature. This protein degrades the collagen found in the extracellular matrix of the stomach mucosa, as well as the collagen involved in ulcer recovery. By delaying the healing process, it thus increases the severity of ulcer. Thus, Core_6 was found to be associated with promoting ulceration.

Acc_20 was found to contain SCOT complex associated with citric acid cycle and \textit{acxC}, a subunit of \textit{acxABC} responsible for acetone carboxylation. While \textit{H. pylori} lacks succinate dehydrogenase, a key enzyme in Kreb’s cycle, the SCOT complex acts as an alternative for energy metabolism. On the other hand \textit{acxABC} promotes acetone utilization. Thus, these two are beneficial for \textit{H. pylori} survival. Mutation is \textit{acxABC} led to reduce the colonization significantly in the stomach of mice [51]. So, Acc_20 plays an important role by providing an alternative path for energy production and assisting \textit{H. pylori} to colonize in gastric mucosa.

LPS biosynthesis genes such \textit{hofC}, \textit{hofD}, and \textit{lpxB} were found to be present as an accessory island. LPS is present in the cellular membrane of all Gram-negative bacteria and plays a crucial role in pathogenesis. LPS has three distinct parts, a core composed of oligosaccharide, lipid A and O antigen. Lipid A is able to trigger fatal immune reaction even in very low concentration, thus known as endotoxin [52]. The core part is found to be conserved in closely related bacteria. The O antigen representing the outer most part of LPS, imitates carbohydrate (glycan) structure found on human epithelial cells by incorporating Lewis antigens on it [53]. The Lewis antigens (Le\textsuperscript{x} and Le\textsuperscript{y}) are capable of interacting with DC-SIGN. It is a type II C-type lectin present on the dendritic cells of gastric epithelium. This
interaction signal down regulates the inflammatory cascade [54]. This suppression of the immune response facilitates a chronic *H. pylori* infection. Thus this GI is also associated with the disease prognosis indirectly. GC is multi factorial disease which includes genetic and environmental factors. The high salt in-taking diet habit was found to be linked with increase the virulence of *H. pylori*. The virulence *cagA* +ve strains were found to be enhanced with heavy salted diet in mice. It elevated the *H. pylori* colonization in the gastric mucosa and led to greater depletion of parietal cell in mice, thus facilitating the development of GC [41]. This result was observed when similar experiment was done with *cagA* -ve strains. So, rich salt diet promotes the GC prognosis in *H. pylori* infected people. Acc_14 was discovered to have a number of genes whose over expression has been linked to a high-salt diet, for example, *amiE, lysA*. The exact function of these proteins in GC prognosis not yet understood.

The colonization potential of *H. pylori* is critically influenced by flagellar motility. The *fliD* mutant strain was completely non-motile, was unable to colonize the stomach mucosa of host mice [55]. The *fliD* operon is linked to flagellar assembly and biofilm formation was found on Core_7. Many bacteria employ flagella to assist swarming, adhesion, and biofilm formation. In biofilms, *H. pylori* may be able to withstand environmental challenges such stomach acidity and ROS produced by phagocytic cells. The bacteria can evade antibiotic therapy by living in a biofilms, leading to chronic infection [37]. Flagellar hook-associated protein 2 (*fliD*) was found to be present across all the *H. pylori* strains and associated with essential function like flagellar biosynthesis and biofilm formation. It can be considered as a good target for development of a new therapeutic to combat with the *H. pylori* infection.

Characterization of accessory GI (Acc_31) led to the identification of a putative labile enterotoxin. Phylogenetic analysis suggested that this toxin was horizontally acquired from another helical-shaped, Gram-negative bacterium *Campylobacter jejuni*, which is the most common cause of food poisoning in Europe and the United States [56]. Further analysis showed that this toxin interacts with ligase *tilS*, an essential protein for viability. *TilS* along with *dus* work in tRNA processing. Apart from this, *tilS* co-express with *ddl* which is involved in the peptidoglycan biosynthesis pathway. Thus, putative toxin-*tilS* interaction indirectly helps in tRNA processing, cell wall biosynthesis and vancomycin resistance. Generally, enterotoxins showed a marked effect in the gastrointestinal tract by increasing the chloride permeability of the intestinal mucosal cells [57-60]. The unique ligand of the putative toxin was sulfate ion, known to inhibit electrolyte absorption and increased intestinal motility. These two are the key regulators of food poisoning. Putative labile enterotoxin output A, a GTPase virulence factor, a dynamin-like protein (DLP) enhances toxin release,
potentially through vesicle secretion [61]. The cellular location of this putative toxin is in the periplasmic region. We can vouch the novel toxin along with other dynamin-like proteins present on the GI involve in vesicle formation, membrane fusion, thus lead to increase toxin release.

Toxin-antitoxin genes are often inherited through HGT [62]. Three Fukui strains (H. pylori F30, H. pylori F16 and H. pylori F32) contain type II TAs which belong to Vap family and encoded by bicistronic operon. VapD was present in 60% of the H. pylori genomes [63]. Design-Island-II identified these TAs in 29 strains. The lengths of the GI in 26 strains were <10Kb, due to deletion of neighboring gene content. Therefore, we studied this GI only in the afore-mentioned strains. Docking of vapD and its antitoxin forms a stable complex with minimum free energy. The antitoxin binds to vapD in the post-translational stage and inhibits its expression. The exact biological role of the vapD protein is yet to be established. Several reports suggest VapD in Haemophilus influenzae, acts as a toxin and in Rhodococcus equi helps in survival through acid tolerance [64, 65]. In adenocarcinoma gastric cells and gastric biopsies, vapD-antitoxin expression values were higher than cagA and vacA cytotoxin genes. Another report showed treatment of biopsy samples with chloramphenicol and kanamycin induced the expression of TAs. These suggested that vapD is a virulent factor, protects the bacterium by aiding biofilm formation [66, 67]. It helps the bacterium to survive in a hostile environment. Thus, the presence of H. pylori in the gastric mucosa for a long time leads to the formation of severe lesions. Integrating all the reports we can presume, under stress and an unfavorable environment the partner anti-toxin represses its own regulation, thus aiding the virulence of vapD.

Various other GI prediction tools were used to validate the identified GIs by the Design-Island-II. This method is an unsupervised method, uses Monte-Carlo statistical test on randomly selected segments of a chromosome. It calculates the oligonucleotide frequencies and performs a statistical test to identify the GIs. Other sequence composition-based approaches, like PredictBias and MSGIP showed comparable prediction with Design-Island-II. In contrary, comparative genomics approach (IslandPick) and hybrid approach (IslandViewer 4) predicted GIs were somewhat different from Design-Island-II predicted results. However, experimental validation of these predicted GIs and their functional implication in disease pathogenesis need to be explored to get a better insight on the role of these GIs.

Finally, we can conclude, along with well-studied cagA and vacA virulent genes, it is equally important to study other virulence factors, as these have direct or indirect impact in the
The chronicity of gastric diseases like, some of them were linked to induce ulceration or bacterial colonization, while others aided in biofilm development, allowing bacteria to elude therapy and promote persistent infection.

**Competing Interests**

The author declares that there are no competing interests associated with the manuscript.

**Acknowledgments**

JC is supported by DST, Govt. of India. The work was supported by Indian Statistical Institute, Kolkata, India. RC is supported by the intramural funding of Indian Statistical Institute and Technology Innovation Hub (TIH) on Data Science, Big Data Analytics, and Data Curation at ISI, funded by DST Govt. of India.

**Data availability**

The data generated in the manuscript are provided in supplementary files. The sources of data used in this manuscript are referred in the manuscript.

**References**

1. Conteduca, V., et al., *H. pylori infection and gastric cancer: state of the art (review)*. Int J Oncol, 2013. 42(1): p. 5-18.
2. Rolig, A.S., et al., *Helicobacter pylori requires TlpD-driven chemotaxis to proliferate in the antrum*. Infect Immun, 2012. 80(10): p. 3713-20.
3. Fock, K.M., D.Y. Graham, and P. Malfertheiner, *Helicobacter pylori research: historical insights and future directions*. Nat Rev Gastroenterol Hepatol, 2013. 10(8): p. 495-500.
4. Watanabe, T., et al., *Helicobacter pylori infection induces gastric cancer in mongolian gerbils*. Gastroenterology, 1998. 115(3): p. 642-8.
5. Covacci, A., et al., *Helicobacter pylori virulence and genetic geography*. Science, 1999. 284(5418): p. 1328-33.
6. Uemura, N., et al., *Helicobacter pylori infection and the development of gastric cancer*. N Engl J Med, 2001. 345(11): p. 784-9.
7. Weinstock, G.M., *Genomics and bacterial pathogenesis*. Emerg Infect Dis, 2000. 6(5): p. 496-504.
8. Juhas, M., et al., *Genomic islands: tools of bacterial horizontal gene transfer and evolution*. FEMS Microbiol Rev, 2009. 33(2): p. 376-93.
9. Moriel, D.G., et al., *Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic Escherichia coli*. Proc Natl Acad Sci U S A, 2010. 107(20): p. 9072-7.
10. Dobrindt, U., et al., *Genomic islands in pathogenic and environmental microorganisms*. Nat Rev Microbiol, 2004. 2(5): p. 414-24.
11. Hudson, C.M., et al., *Resistance determinants and mobile genetic elements of an NDM-1-encoding Klebsiella pneumoniae strain*. PLoS One, 2014. 9(6): p. e99209.
12. Kumar, S., et al., *MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences*. Brief Bioinform, 2008. 9(4): p. 299-306.

13. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated on the tree of life*. Nucleic Acids Res, 2015. 43(Database issue): p. D447-52.

14. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis*. Nat Protoc, 2015. 10(6): p. 845-58.

15. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography*. Acta Crystallogr D Biol Crystallogr, 2010. 66(Pt 1): p. 12-21.

16. Schmidtke, P., et al., *fpocket: online tools for protein ensemble pocket detection and tracking*. Nucleic Acids Res, 2010. 38(Web Server issue): p. W582-9.

17. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data*. Nucleic Acids Res, 2004. 32(Database issue): p. D129-33.

18. DeLano, W.L. *Pymol: An open-source molecular graphics tool*. 2002.

19. Kozakov, D., et al., *The ClusPro web server for protein-protein docking*. Nat Protoc, 2017. 12(2): p. 255-278.

20. Sanchez-Garcia, R., et al., *BIPSPI: a method for the prediction of partner-specific protein-protein interfaces*. bioinformatics, 2019. 35(3): p. 470-477.

21. Pundhir, S., H. Vijayvargiya, and A. Kumar, *PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes*. In Silico Biol, 2008. 8(3-4): p. 223-34.

22. Wei, W., et al., *Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties*. Brief Bioinform, 2017. 18(3): p. 357-366.

23. de Brito, D.M., et al., *A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm*. PLoS One, 2016. 11(1): p. e0146352.

24. Langille, M.G., W.W. Hsiao, and F.S. Brinkman, *Evaluation of genomic island predictors using a comparative genomics approach*. BMC Bioinformatics, 2008. 9: p. 329.

25. Bertelli, C., et al., *IslandViewer 4: expanded prediction of genomic islands for large-scale datasets*. Nucleic Acids Res, 2017. 45(W1): p. W30-W35.

26. Chatterjee, R., K. Chaudhuri, and P. Chaudhuri, *On detection and assessment of statistical significance of Genomic Islands*. BMC Genomics, 2008. 9: p. 150.

27. Nomura, M., *Engineering of bacterial ribosomes: replacement of all seven Escherichia coli rRNA operons by a single plasmid-encoded operon*. Proc Natl Acad Sci U S A, 1999. 96(5): p. 1820-2.

28. Yap, W.H., Z. Zhang, and Y. Wang, *Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon*. J Bacteriol, 1999. 181(17): p. 5201-9.

29. Tsirigos, A. and I. Rigoutsos, *A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes*. Nucleic Acids Res, 2005. 33(12): p. 3699-707.

30. Garcia-Vallve, S., et al., *HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes*. Nucleic Acids Res, 2003. 31(1): p. 187-9.

31. Vinogradov, A.E., *Isochores and tissue-specificity*. Nucleic Acids Res, 2003. 31(17): p. 5212-20.

32. Wei, K., T. Zhang, and L. Ma, *Divergent and convergent evolution of housekeeping genes in human-pig lineage*. PeerJ, 2018. 6: p. e4840.

33. Labigne, A., *Shuttle cloning and nucleotide sequences of Helicobacter pylori genes responsible for urease activity*. J Bacteriol, 1991.
34. Cussac, V., Expression of Helicobacter pylori urease genes in Escherichia coil grown under nitrogen-limiting conditions. J Bacteriol, 1992.
35. Gillessen, A., et al., Evidence of de novo collagen synthesis in healing human gastric ulcers. Scand J Gastroenterol, 1995. 30(6): p. 515-8.
36. Ratthawongjirakul, P., The impacts of a fltD mutation on the biofilm formation of Helicobacter pylori. Asian Pacific Journal of Tropical Biomedicine, 2016. 6(12): p. 1008-1014.
37. Garcia, A., et al., Biofilm and Helicobacter pylori: from environment to human host. World J Gastroenterol, 2014. 20(19): p. 5632-8.
38. Hug, I., et al., Helicobacter pylori lipopolysaccharide is synthesized via a novel pathway with an evolutionary connection to protein N-glycosylation. PLoS Pathog, 2010. 6(3): p. e1000819.
39. Tsugane, S., Salt, salted food intake, and risk of gastric cancer: epidemiologic evidence. Cancer Sci, 2005. 96(1): p. 1-6.
40. Fox, J.G., et al., Helicobacter pylori-associated gastric cancer in INS-GAS mice is gender specific. Cancer Res, 2003. 63(5): p. 942-50.
41. Fox, J.G., et al., High-salt diet induces gastric epithelial hyperplasia and parietal cell loss, and enhances Helicobacter pylori colonization in C57BL/6 mice. Cancer Res, 1999. 59(19): p. 4823-8.
42. Cover, T.L. and R.M. Peek, Jr., Diet, microbial virulence, and Helicobacter pylori-induced gastric cancer. Gut Microbes, 2013. 4(6): p. 482-93.
43. Voss, B.J., et al., Alteration of the Helicobacter pylori membrane proteome in response to changes in environmental salt concentration. Proteomics Clin Appl, 2015. 9(11-12): p. 1021-34.
44. Mills, S.D., W. Yang, and K. MacCormack, Molecular characterization of benzimidazole resistance in Helicobacter pylori. Antimicrob Agents Chemother, 2004. 48(7): p. 2524-30.
45. Debowski, A.W., et al., Helicobacter pylori gene silencing in vivo demonstrates urease is essential for chronic infection. PLoS Pathog, 2017. 13(6): p. e1006464.
46. Athmann, C., et al., Local pH elevation mediated by the intrabacterial urease of Helicobacter pylori coccultured with gastric cells. J Clin Invest, 2000. 106(3): p. 339-47.
47. Weeks, D.L., et al., A H+-gated urea channel: the link between Helicobacter pylori urease and gastric colonization. Science, 2000. 287(5452): p. 482-5.
48. Censini, S., 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(25): p. 14648-53.
49. Wiedemann, T., et al., Helicobacter pylori cag-Pathogenicity island-dependent early immunological response triggers later precancerous gastric changes in Mongolian gerbils. PLoS One, 2009. 4(3): p. e4754.
50. Ogura, K., et al., Virulence factors of Helicobacter pylori responsible for gastric diseases in Mongolian gerbil. J Exp Med, 2000. 192(11): p. 1601-10.
51. Brahmacary, P., et al., The human gastric pathogen Helicobacter pylori has a potential acetone carboxylase that enhances its ability to colonize mice. BMC Microbiol, 2008. 8: p. 14.
52. Raetz, C.R. and C. Whitfield, Lipopolysaccharide endotoxins. Annu Rev Biochem, 2002. 71: p. 635-700.
53. Moran, A.P., Relevance of fucosylation and Lewis antigen expression in the bacterial gastroduodenal pathogen Helicobacter pylori. Carbohydr Res, 2008. 343(12): p. 1952-65.
54. Bergman, M.P., et al., *Helicobacter pylori modulates the T helper cell 1/T helper cell 2 balance through phase-variable interaction between lipopolysaccharide and DC-SIGN*. J Exp Med, 2004. 200(8): p. 979-90.

55. Kim, J.S., et al., *Molecular cloning and characterization of the Helicobacter pylori fliD gene, an essential factor in flagellar structure and motility*. J Bacteriol, 1999. 181(22): p. 6969-76.

56. Silva, J., et al., *Campylobacter spp. as a Foodborne Pathogen: A Review*. Front Microbiol, 2011. 2: p. 200.

57. Lucas, M.L., *Amendments to the theory underlying Ussing chamber data of chloride ion secretion after bacterial enterotoxin exposure*. J Theor Biol, 2005. 234(1): p. 21-37.

58. Lucas, M.L., *Enterocyte chloride and water secretion into the small intestine after enterotoxin challenge: unifying hypothesis or intellectual dead end?* J Physiol Biochem, 2008. 64(1): p. 69-88.

59. Pinchuk, I.V., E.J. Beswick, and V.E. Reyes, *Staphylococcal enterotoxins*. Toxins (Basel), 2010. 2(8): p. 2177-97.

60. Principato, M. and B.F. Qian, *Staphylococcal enterotoxins in the etiopathogenesis of mucosal autoimmunity within the gastrointestinal tract*. Toxins (Basel), 2014. 6(5): p. 1471-89.

61. Michie, K.A., et al., *LeoA, B and C from enterotoxigenic Escherichia coli (ETEC) are bacterial dynamin*. PLoS One, 2014. 9(9): p. e107211.

62. Ramisetty, B.C., et al., *What Is the Link between Stringent Response, Endoribonuclease Encoding Type II Toxin-Antitoxin Systems and Persistence?* Front Microbiol, 2016. 7: p. 1882.

63. Cao, P. and T.L. Cover, *High-level genetic diversity in the vapD chromosomal region of Helicobacter pylori*. J Bacteriol, 1997. 179(9): p. 2852-6.

64. Benoit, S., et al., *Induction of vap genes encoded by the virulence plasmid of Rhodococcus equi during acid tolerance response*. Res Microbiol, 2001. 152(5): p. 439-49.

65. Daines, D.A., J. Jarisch, and A.L. Smith, *Identification and characterization of a nontypeable Haemophilus influenzae putative toxin-antitoxin locus*. BMC Microbiol, 2004. 4: p. 30.

66. Cardenas-Mondragon, M.G., et al., *Transcriptional Profiling of Type II Toxin-Antitoxin Genes of Helicobacter pylori under Different Environmental Conditions: Identification of HP0967-HP0968 System*. Front Microbiol, 2016. 7: p. 1872.

67. Morales-Espinosa, R., et al., *High expression of Helicobacter pylori VapD in both the intracellular environment and biopsies from gastric patients with severity*. PLoS One, 2020. 15(3): p. e0230220.
Figure Legends

Figure 1. GI distribution in 50 *H. pylori* strains. A. Distribution of forty-one statistically significant GIs in 50 *H. pylori* strains. B. Heatmap with the Kendall Tau's correlation coefficients considering presence or absence of accessory islands among different *H pylori* strains. C. Distribution of islands on the basis of $Sm$ score.

Figure 2. Comparative genomic study of Niche Adaptation Island (Core_5) in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Red: $S_m \geq 0.70$; Grey: $S_m: >0.4-0.69$, Green: $S_m \leq 0.39$. Fifty strains showed three main clusters. B. Synteny of representative strains from each cluster. The region highlighted in red is the most conserved part consists of seven genes in the Urease Island and the variable region is highlighted in green.
**Figure 3.** Comparative genomic study of Lipopolysaccharide Biosynthesis Island (Acc_17) in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Red: $S_n \geq 0.70$; Grey: $0.4-0.69$; Green: $S_n \leq 0.39$. Fifty strains showed one major cluster. B. Synteny of representative strain from main cluster and other strains.

**Figure 4.** Identification and comparative study of putative toxin in *Helicobacter pylori*. A. Synteny of five strains showing to have putative enterotoxin. The putative gene highlighted in red is the enterotoxin. B. Synteny of other strains of the same region. C. Heatmap of similarity scores using hierarchical clustering. Red: $S_n \geq 0.70$; Grey: $0.4-0.69$; Green: $S_n \leq 0.39$. Five strains showed similar structural composition. The region highlighted in blue composed of the strains, having putative enterotoxin.

**Figure 5.** Evolutionary and functional analysis of novel toxin. A. Phylogenetic tree of homologs of putative toxin in different bacteria. B. Conservation of residues in the binding site. Yellow highlighted amino acids are conserved in all the studied microorganisms. C. Protein interaction study of putative toxin. Gene neighborhood and gene co-occurrence are shown in green and blue color respectively. D. Protein 3D homology model of putative toxin. Green highlighted residues are involved in binding pocket formation. E. Zoomed view of the binding pocket.

**Figure 6.** Characterization of TA system. A. Synteny study of the representative sequence of the cluster. B. The interacting residues in vapD-antitoxin docked complex. C. Surface structure of the docked protein. D and E. Residues responsible for protein-protein interaction between vapD and antitoxin respectively. Red: vapD, Green: Interacting residues of vapD, Blue: DUF3240 family anti-toxin, Orange: Interacting residues of DUF3240 family anti-toxin.

**Supplementary Figure Legends**

**Supplementary Figure 1.** Comparative genomic study of Core_3 in fifty strains of *Helicobacter pylori*. A. Heatmap of Core_3 generated from square matrices of similarity score among the fifty *H. pylori* strains. B. Synteny of Core_3 Genomic island in all fifty *H. pylori* strains. Four missense mutations in the NuoB and NuoD genes have resulted in the emergence of benzimidazole-resistant *H. pylori* strains.

**Supplementary Figure 2.** Comparative genome study of Core_4. A. Synteny of representative strains of Core_4. B. Heatmap of similarity scores using hierarchical
clustering. Fifty strains showed two major clusters. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$.

**Supplementary Figure 3.** Comparative genomic study of Core_6 in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Fifty strains showed one main cluster. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$. B. Synteny of representative strains of Core_6. U32 family protein was conserved among the strains.

**Supplementary Figure 4.** Comparative genomic study of Core_7 in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Fifty strains formed three major clusters. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$. Fifty strains showed two main clusters. B. Synteny of representative strains from each clusters. *fliD* is present in all the fifty strains.

**Supplementary Figure 5.** Comparative genomic study of Cag Pathogenicity Island (Acc_25) in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$. Fifty strains showed four main clusters. Cluster4 represents cag –ve strains. B. Synteny of representative strains from each cluster. The region highlighted in red is the most variable part of the Cag PAI.

**Supplementary Figure 6.** Comparative genomic study of Acc_14 in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$. B. Synteny of representative strains based on the presence and absence of *lysA* and *amiE* gene.

**Supplementary Figure 7.** Comparative genomic study of Acc_20 in fifty strains of *Helicobacter pylori*. A. Heatmap of similarity scores using hierarchical clustering. Fifty strains showed one main cluster. Red: $S_m \geq 0.70$; Grey: $S_m: \geq 0.4$-0.69, Green: $S_m \leq 0.39$. B. Synteny of representative strains from each cluster. The region highlighted in red is the most conserved part.

**Supplementary Figure 8.** Quality assessment of homology model of putative toxin. A. Rotamer analysis: Some side chains in the model may not have been modelled ideally. These are coloured in red and may indicate a problem with the backbone or underlying alignment in this region. B. Ramachandran analysis: Some residues in the model may lie in favorable (blue), allowed (green) or disallowed (red) regions of the Ramachandran plot. This colouring indicates residues that may have problems with the backbone phi/psi angles. C. Clash
Analysis: Some atoms in some residues may lie too close to one another in the model. Residues are colored by how many clashes are observed. A large number of clashes could mean bad side chain placement.

**Supplementary Figure 9:** Synteny study of the Acc_33. The representative synteny of eight *H. pylori* strains which were not identified by PredictBias is highlighted in the figure.
Toxin-Antitoxin System

H. pylori 26695, Pecan4, Pecan18, P12

H. pylori Cuz20, v225D, Puno120, 52, BM012A, XZ274

H. pylori F16, F30, OK113, Shi112, Sat464, SNT49
  OK310, B8, SA7, SA20, F32

H. pylori Aklavik117, Rif1, Shi417, Um037

H. pylori Puno135, India7

H. pylori Shi169, Rif2

 Detected by Design-Island-II, MSGIP and PredictBias

 Detected by Design-Island-II and MSGIP
### Supplementary Table 1: Disease association of 50 *H. pylori* strains

| Disease               | Number of strains |
|-----------------------|-------------------|
| Duodenal Ulcer (DU)   | 8                 |
| Gastric Cancer (GC)   | 6                 |
| Gastric Ulcer (GU)    | 3                 |
| Gastritis             | 7                 |
| Peptic Ulcer (PU)     | 3                 |
| Unknown               | 23                |

### Supplementary Table 2: Geographical Distribution of 50 *H. pylori* strains

| Country                          | Number of strains |
|----------------------------------|-------------------|
| Canada                           | 2                 |
| Malaysia                         | 3                 |
| UK                               | 1                 |
| Sweden                           | 1                 |
| Spain                            | 1                 |
| Lithuania                        | 1                 |
| Italy                            | 1                 |
| Germany                          | 1                 |
| France                           | 2                 |
| Korea                            | 2                 |
| Japan                            | 8                 |
| China                            | 1                 |
| Australia                        | 2                 |
| India                            | 2                 |
| Venezuela                        | 1                 |
| Peru                             | 14                |
| West and South African Countries | 7                 |
**Supplementary Table 3**: Functional Characterization of thirty eight accessory islands identified by comparative study fifty *H. pylori* strains

| Name       | Function                          | Name    | Function                          |
|------------|-----------------------------------|---------|-----------------------------------|
| Core_1     | Ribosomal Protein                 | Acc_14  | Lysine biosynthetic process       |
| Core_2     | Ribosomal Protein                 | Acc_15  | Miscellaneous                     |
| Core_3     | Antibiotic Resistance Island      | Acc_16  | Miscellaneous                     |
| Core_4     | VacA island                       | Acc_17  | Lipopolysaccharide biosynthesis   |
| Core_5     | Niche adaptation                  | Acc_18  | Miscellaneous                     |
| Core_6     | Promoting Ulceration              | Acc_19  | Miscellaneous                     |
| Core_7     | Biofilm Formation                 | Acc_20  | Metabolism                        |
| Acc_1      | Niche adaptation                  | Acc_21  | Miscellaneous                     |
| Acc_2      | Miscellaneous                      | Acc_22  | Miscellaneous                     |
| Acc_3      | Miscellaneous                      | Acc_23  | Membrane efflux protein           |
| Acc_4      | Bacterial Transport System         | Acc_24  | Miscellaneous                     |
| Acc_5      | Biosynthesis                      | Acc_25  | Cag pathogenicity Island          |
| Acc_6      | Miscellaneous                      | Acc_26  | Miscellaneous                     |
| Acc_7      | Miscellaneous                      | Acc_27  | Metabolism                        |
| Acc_8      | Miscellaneous                      | Acc_28  | Terpenoid-quinone biosynthesis    |
| Acc_9      | Miscellaneous                      | Acc_29  | Cell cycle                        |
| Acc_10     | Miscellaneous                      | Acc_30  | Metabolism                        |
| Acc_11     | Miscellaneous                      | **Acc_31** | **Putative toxin**              |
| Acc_12     | Miscellaneous                      | Acc_32  | Purine biosynthesis               |
| Acc_13     | Miscellaneous                      | **Acc_33** | **Toxin- antitoxin system**      |
| Acc_34     |                                    |         | Miscellaneous                     |
**Supplementary Table 4**: Number of *H. pylori* strains detected by different GI prediction tools

| GI Names | Design-Island-II | MSGIP | PredictBias | ZislandExplorer | IslandPath-DIMOB | IslandPick |
|----------|------------------|-------|--------------|-----------------|------------------|------------|
| Acc_1    | 32               | 0     | 0            | 0               | 0                | 0          |
| Acc_2    | 26               | 26    | 26           | 0               | 0                | 0          |
| Acc_3    | 34               | 26    | 31           | 25              | 0                | 0          |
| Acc_4    | 49               | 0     | 0            | 0               | 0                | 0          |
| Acc_5    | 26               | 0     | 0            | 0               | 0                | 0          |
| Acc_6    | 27               | 27    | 27           | 27              | 0                | 0          |
| Acc_7    | 26               | 16    | 26           | 0               | 0                | 0          |
| Acc_8    | 39               | 34    | 37           | 31              | 0                | 0          |
| Acc_9    | 45               | 0     | 0            | 0               | 0                | 0          |
| Acc_10   | 46               | 46    | 0            | 0               | 0                | 0          |
| Acc_11   | 45               | 40    | 38           | 0               | 0                | 0          |
| Acc_12   | 33               | 31    | 29           | 35              | 0                | 0          |
| Acc_13   | 32               | 32    | 19           | 36              | 0                | 0          |
| Acc_14   | 33               | 0     | 0            | 0               | 0                | 0          |
| Acc_15   | 39               | 32    | 32           | 36              | 0                | 0          |
| Acc_16   | 27               | 22    | 26           | 24              | 0                | 0          |
| Acc_17   | 41               | 0     | 0            | 0               | 0                | 0          |
| Acc_18   | 39               | 32    | 33           | 0               | 0                | 0          |
| Acc_19   | 44               | 42    | 37           | 35              | 0                | 0          |
| Acc_20   | 48               | 0     | 0            | 0               | 0                | 0          |
| Acc_21   | 47               | 43    | 42           | 40              | 0                | 0          |
| Acc_22   | 32               | 34    | 0            | 0               | 0                | 0          |
| Acc_23   | 41               | 41    | 34           | 0               | 0                | 0          |
| Acc_24   | 44               | 42    | 0            | 0               | 0                | 0          |
| Acc_25   | 45               | 45    | 45           | 45              | 45               | 45         |
| Acc_26   | 45               | 0     | 0            | 0               | 0                | 0          |
| Acc_27   | 40               | 0     | 0            | 0               | 0                | 0          |
| Acc_28   | 43               | 0     | 0            | 0               | 0                | 0          |
| Acc_29   | 30               | 30    | 26           | 26              | 0                | 0          |
| Acc_30   | 30               | 30    | 27           | 24              | 0                | 0          |
| **Acc_31** | **5**         | **5** | **5**        | **5**          | **0**            | **0**     |
| **Acc_32** | **35**       | **35** | **0**        | **0**          | **0**            | **0**     |
| **Acc_33** | **29**     | **29** | **21**       | **0**          | **0**            | **0**     |
| **Acc_34** | **10**   | **0**  | **0**        | **0**          | **0**            | **0**     |