BMC Cancer

RESEARCH

Open Access

Comparative genomic analysis of *Helicobacter pylori* isolates from gastric cancer and gastritis in China

Peng-fei Kong[1†], Yong-hao Yan[1†], Yan-tao Duan[1†], Yan-tian Fang[1†], Yi Dou[1], Yong-hu Xu[1] and Da-zhi Xu[1*]

Abstract

Background This study aimed to explore and compare the genomic characteristics and pathogenicity of *Helicobacter pylori* (*H. pylori*) strains derived from the gastric cancer (GC) and gastritis in the Chinese population.

Methods We performed whole genome sequencing on 12 *H. pylori* strains obtained from GC and gastritis patients in China. Additionally, we retrieved sequencing data for 20 *H. pylori* strains from various regions worldwide from public databases to serve as reference genomes. An evolutionary tree was constructed based on comparative genomics, and we analyzed the differences in virulence factors (VFs) and gene functions.

Results In the GC strains, we identified 1,544 to 1,640 coding genes, with a total length ranging from 1,549,790 to 1,605,249 bp. In the gastritis strains, we found 1,552 to 1,668 coding genes, with a total length spanning from 1,552,426 to 1,665,981 bp. The average length of coding genes was approximately 1,594 (90.91%) for GC strains and 1,589 (90.81%) for gastritis strains. We observed a high degree of consistency in the VFs predicted for both cohorts; however, there was a significant difference in their *cagA* status. Clustering analysis showed significant core single nucleotide polymorphisms (SNPs) differences between GC and gastritis strains, but no major differences in homologous proteins or gene islands. Subsequent pan-genomic and Average Nucleotide Identity (ANI) analyses indicated high homology among GC, gastritis, and other reference *H. pylori* strains. Furthermore, gene function annotation results showed substantial similarity in gene functions between the *H. pylori* strains from GC and gastritis patients, with specific functions primarily concentrated in metabolic processes, transcription, and DNA repair.

Conclusions *H. pylori* strains derived from GC and gastritis patients exhibit differences in virulence factors and SNPs, yet they demonstrate high genomic homology across other levels in the Chinese population.

Keywords Gastric cancer, Gastritis, *Helicobacter pylori*, Genome

[†]Peng-fei Kong, Yong-hao Yan, Yan-tao Duan and Yan-tian Fang contributed equally to this work.

*Correspondence:
Da-zhi Xu
xudzh@shca.org.cn
[1]Department of Gastric Surgery, Precision Cancer Medicine Center, Fudan University Shanghai Cancer Center, 270 Dong'an Road, Shanghai 200032, China

Background

Helicobacter pylori (*H. pylori*) is a highly successful gram-negative bacterium that chronically infects the human gastric mucosa, affecting an estimated 4.3 billion people worldwide in 2022 [1]. *H. pylori* infection is linked to gastritis and gastric cancer (GC), with approximately 1–3% of infected individuals eventually developing GC [2, 3]. The prevalence of *H. pylori* infection is notably higher in developing countries compared to developed nations [4]. Currently, China has one of the highest infection rates worldwide, exceeding 50% [5]. Meanwhile, according to GLOBOCAN 2022, the age-standardized rate (ASR) of GC prevalence in Eastern Asia is alarmingly high at 16.1% [6]. However, there is minimal knowledge about the origin of *H. pylori* genomes and pathogenicity differences in patients with gastritis and GC in this region, which limits the application of molecular diagnostics, such as screening for early GC, prevention of precancerous lesions, and vaccine drug development.

Previously, several studies applied the first-generation sequencing method to initially explore and compare the genetic differences of *H. pylori* from GC and benign gastric diseases [7–9]. In 2009, by comparing *H. pylori* from GC and benign gastric ulcers, it has been first found that highly differentiated alleles and strain specific genes may be useful biomarkers for analyzing the geographic distribution of *H. pylori* and identifying strains capable of inducing malignant or precancerous gastric lesions [10]. Following this, subsequent research has increasingly focused on specific genes and virulence factors (VFs), such as the *vacA* and *cagA* genes, which may serve as biomarkers for GC risk [11, 12]. Despite the utilization of whole-genome sequencing (WGS) in some studies, there is still a significant gap in research regarding the specific genomic differences between *H. pylori* strains in GC and gastritis patients, particularly regarding the genetic variations of H. pylori strains within the Chinese population. In this study, we isolated and cultured *H. pylori* strains from patients with GC and gastritis (GC: $n = 6$, gastritis: $n = 6$), and employed whole-genome sequencing methods to thoroughly explore and compare the genomic and pathogenic differences between these two cohorts in the Chinese population.

Methods

H. Pylori isolates

Twelve strains of *H. pylori* were isolated from tissue samples of patients undergoing gastric surgery (open surgery and endoscopic resection) at Fudan University Shanghai Cancer Center (FUSCC) (Supplemental Table S1). Based on the diagnosis, the isolates were categorized into two subgroups: GC ($n = 6$) and gastritis ($n = 6$). With the consent of all patients, at least 2 pieces of tissue were collected from the margin of either benign lesions or cancerous tissue for bacterial culture and pathological examination. Two experienced clinical pathologists performed pathological assessments of the gastric tissue specimens. This study was approved by the FUSCC review board in accordance with Chinese bioethical regulations, and written informed consent was obtained from all participants (Ethics Approval Number: FUSCC-D-2022-144). In addition, genome sequences of 20 *H. pylori* reference strains were downloaded from GenBank, with detailed information provided in Supplemental Table S2.

Genomic DNA extraction and sequencing

The DNA of *H. pylori* isolates were extracted using a bacterial genomic DNA extraction kit (Beyotime Biotechnology Co., Ltd., China) according to the manufacturer's instructions [13]. The DNA content and purity were assessed using a Qubit fluorometer and a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Carlsbad, USA). WGS of the H. pylori strains was performed at Shanghai PersonalBio Technology Company (SPTC, Shanghai) using the Illumina NovaSeq 6000 platform.

Data preparation and genome assembly

Raw sequencing data were filtered using fastp (version 0.20.0, https://github.com/OpenGene/fastp) to remove sequencing junctions and primers from the reads. Additionally, reads with a constant quality value ≤ 5 bases and those containing more than 5% N bases were discarded, along with any repeated contaminants, resulting in the acquisition of clean reads [14]. Subsequently, the clean data were assembled using the default parameters of Unicycler (v0.4.8), leading to the generation of the genome sequence [15].

Genomic structural and functional annotation

Structural annotation of the genome sequence was performed using prokka(v 1.12) software. Following the generation of the gene set for the sequenced strains, the genes were compared and annotated across multiple databases to determine their functions and related descriptions. The amino acid sequences of the genes from each strain were matched with entries in each database to obtain corresponding functional annotation information. For general functional analysis, three key databases were selected: the Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Groups (COG), and Gene Ontology (GO). Additionally, the Pathogen Virulence Factor Database (VFDB) was utilized to assess pathogenicity.

Comparative genomic analysis

OrthoFinder (v2.3.5) software was used to compare the predicted protein sequences of 32 strains of *H. pylori* and identify homologous gene clusters.

Then, the basic situation of core and non-core genes was analyzed, and the differences within species were studied from the point of view of specific gene sequences [16]. The single-copy homologous genes identified by OrthoFinder were aligned using MAFFT (v 7.427). To identify core SNPs, Snippy (v4.6.0) and Gubbins (v2.4.1) were utilized, and a phylogenetic tree was constructed using FastTree (v2.1.10) and iTOL. Average Nucleotide Identity (ANI) is an index that compares the relationship between two genomes at the nucleotide level, calculating their similarity or evolutionary distance for species classification and kinship comparison. ANI analysis was conducted on the 32 sorted strains in pairwise comparisons using FastANI (v1.31), and the results were organized into a matrix for visualization. Pan-genome is a general term for all genes of a species, reflecting all genetic information at the gene level of the species [17].

Statistical analyses

Statistical analyses were performed using SPSS software version 19.0. Differences between the groups were compared using Chi-square or Pearson Chi-square tests, Fisher's exact method, or Kruskal-Wallis analysis according to the data type and the number of comparisons. Statistical significance were considered clinically significant at a p value < 0.05. All P-values were two-sided. When the results of Pearson Chi-square tests showed a statistical difference, the intra-group pairwise comparison was adjusted using the Bonferroni formula.

Results

Clinical characteristics of patient information and strains isolation

Between 1 March 2021 and 31 October 2022, thirty-nine GC and gastritis patients were hospitalized in our department for surgery and twenty patients tested with *H. pylori* positive. Next, *H. pylori* isolate from each patient (from ten single colony isolates stored for each patient) was selected and sub-cultured. Finally, twelve *H. pylori* isolates were chosen through a three-step randomization procedure for WGS analysis (Fig. 1). Of these patients, six underwent gastrectomy for GC and six underwent Endoscopic submucosal dissection (ESD) for gastritis, respectively. The average age of the patients was 46.2 years, with a range of 37 to 69 years. 75% (8/12) of the patients were male (Supplemental Table S1).

The whole genome of *H. pylori* strains were sequenced and analyzed using Illumina NovaSeq sequencing platforms in this study. Table 1 presents the general characteristics for each of the 12 genome sequences. The number of contigs ranged from 24 to 43, with a high (310×-1040×) genome coverage. The mean genome size was 1.57 Mb, and the average G + C content was 38.8%. Each genome contains between 1544 and 1688 annotated average coding genes sequence (CDS), with approximately 91% of the genome allocated to coding regions.

Circle diagram of the chromosome of *H. Pylori* isolates strain

The genome sequence of *H. pylori* in GC strains consisted of a circular chromosome with an average total length of 1,579,639 bp and an average G + C content of 38.8% (Table 1; Fig. 2A). In the six GC stains of *H. pylori*, between 1544 and 1640 coding genes were predicted, with a total length ranging from 1,549,790 to 1,605,249 bp, accounting for 90.63–91.16% of the total genome. In the gastritis stains of *H. pylori* (Fig. 2B), the 1552–1668 coding genes, with a total length of 1,552,426-1,665,981 bp, were predicted in the genome, which accounted for 90.14-91.19% of the total genome. In addition, the average length of coding genes in GC and gastritis strain, was approximately 1594 (90.91%) and 1589 genes (90.81%), respectively. Based on the VFDB (Virulence Factors of Pathogenic Bacteria) database, these virulence genes were divided into 7 VF (virulence factor) classes including acid resistance, adherence, immune evasion, immune modulator, motility and so forth. In detail, each VF class contained several kinds of VFs and the related genes as shown in Table 2. We compared the VFs of *H. pylori* derived from GC and gastritis, and found that the VFs predicted by the two cohort strains had a high consistency, but there were also obvious differences. For example, *cagA*, a well-known virulence factor related gene, was found in GC strains compared to those derived from gastritis.

Phylogenetic tree of the 12 *H. pylori* isolates and the reference genomes

The genomes of our *H. pylori* isolates were analyzed for phylogenetic relationships using core genome single nucleotide polymorphisms (SNPs) in conjunction with 20 public reference genomes (Supplemental Table S1). The core genome analysis revealed four distinct populations based on geographic origin. Specifically, 30.0% (6/20) of the isolates were from Asia, 30.0% (6/20) from America, 30.0% (6/20) from Europe, and 10.0% (2/20) from Australia.

The phylogenomics of our *H. pylori* strains were analyzed following the flow diagram outlined in Supplementary Figure S1. Pangenome analysis using OrthoFinder identified a total of 2,492 genes across 32 *H. pylori* strains, of which 1,054 genes (42.4%) were conserved and present in all strains (100%) (Supplementary Table S3). We used snippy to identify SNPs of all species using CP0

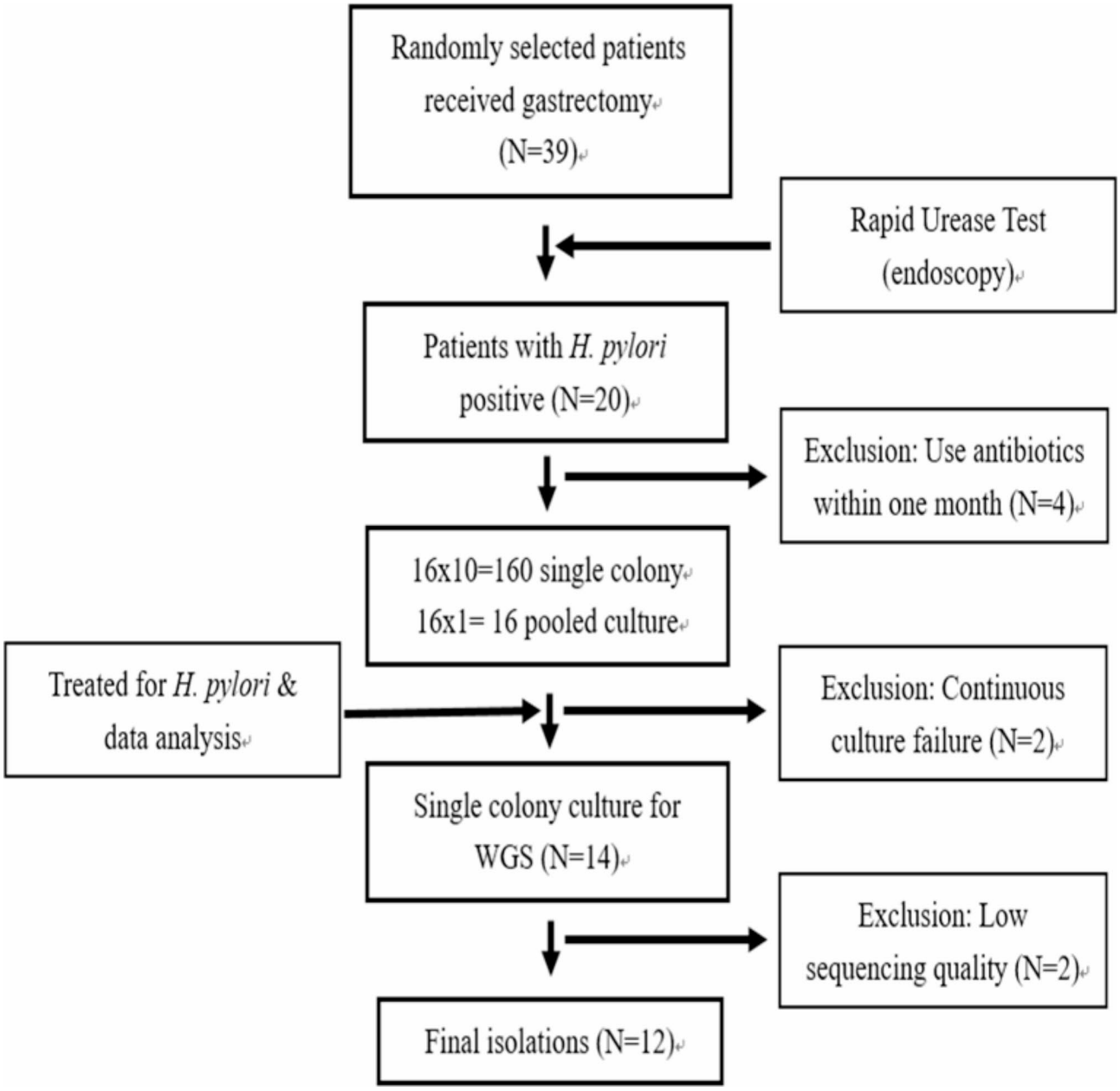

Fig. 1 *H. pylori* isolation, culture in this study and selection of strains for the present study. WGS, Whole Genome Sequencing

03904.1_*Helicobacter_pylori*_26695 (1,667,892 bp) as a reference sequence. A total of 206,091 core SNPs were identified. Gubbins were then used to reassemble 188,039 SNPs, and Fasttree was subsequently used to construct core SNP evolutionary tree. Additionally, ModelFinder was applied to select the best-fit substitution model based on the grouping relationships and cross-validation errors, resulting in the phylogenetic tree depicted in Supplementary Figure S2. The phylogenetic analysis of 20 strains, as illustrated in Fig. 3 and Supplementary Figure S3, revealed four distinct H. pylori populations: HP-Asia (Japan-1, Japan-2, India-1, India-2, Bangladesh-1, and Bangladesh-2), HP-America (United states-1, United states-2, Mexico-1, Mexico-2, Brazil-1, and Brazil-2), HP-Europe (France-1, France-2, Belgium-1, Belgium-2, Spain-1, and Spain-2), and HP-Oceania (Australia-1, and Australia-2). According to the clustering with reference strains, our strains (GC and gastritis strains) were assigned to HP-Asia population, especially with high homology with Japanese strains (Fig. 3A). Furthermore, we also mapped the evolutionary tree of strains based on homologous proteins, corroborating the high similarity of our strains with those from Asia, especially Japanese strains (Fig. 3B). Notably, the clustering results indicated

Table 1 Genome statistics of the whole genome sequences of the 12 *H. pylori* isolates

Strain ID	Genome coverage	No. of contigs	Genome size	No. of CDS	Coding percentage	G+C percentage	rRNA	tRNA
Gastritis-01	830	35	1,540,791	1552	90.89	38.91	2	36
Gastritis-02	1040	24	1,538,881	1553	91.19	38.89	2	36
Gastritis – 03	310	43	1,665,981	1688	90.14	38.61	2	36
Gastritis-04	1020	36	1,563,078	1578	90.92	38.80	2	36
Gastritis-05	800	36	1,555,888	1598	90.57	38.73	2	36
Gastritis-06	610	36	1,552,426	1565	91.17	38.86	2	36
GC-01	580	27	1,591,503	1607	91.09	38.72	2	36
GC-02	760	30	1,549,790	1544	91.16	38.85	2	36
GC-03	960	40	1,560,160	1590	90.63	38.88	2	36
GC-04	490	40	1,573,950	1598	90.90	38.88	2	36
GC-05	1000	28	1,597,182	1583	90.80	38.75	2	36
GC-06	800	33	1,605,249	1640	90.87	38.70	2	36

CDS, Coding DNA Sequence. GC, Gastric Cancer

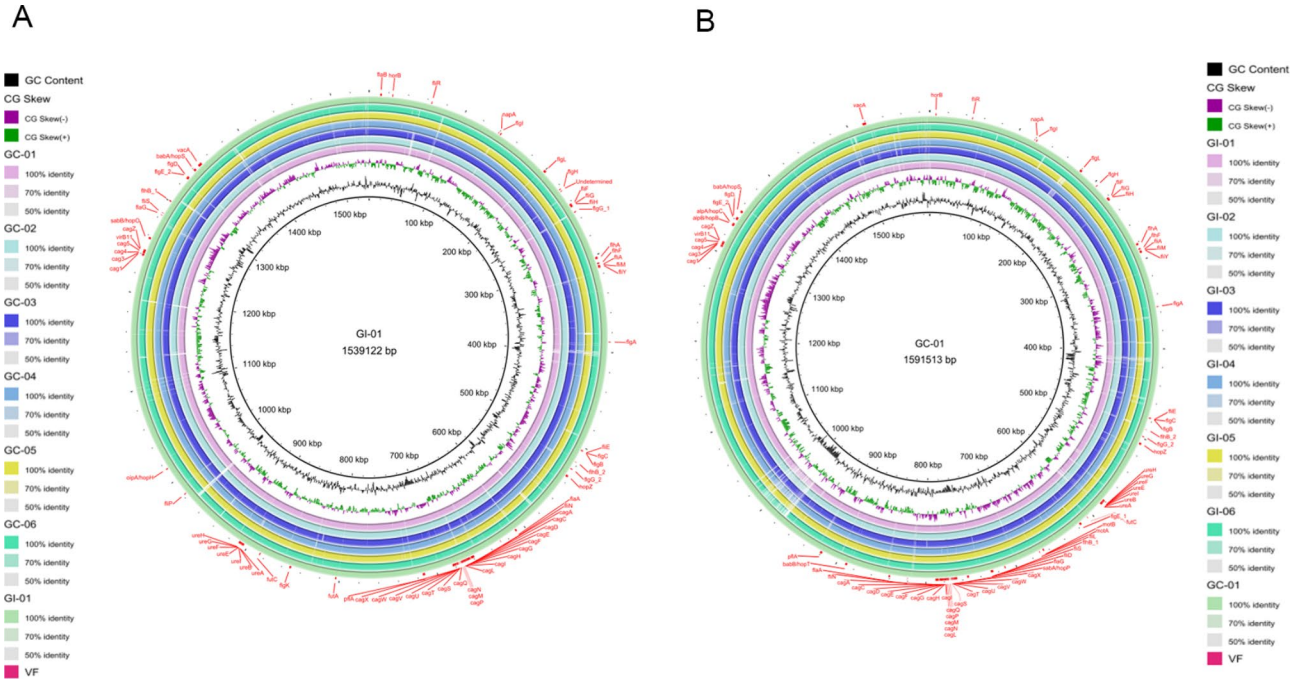

Fig. 2 Circle diagram of the chromosome of*H. pylori*isolates strain in GC and gastritis patients. (**A**) Circle diagram of the chromosome of *H. pylori* isolates strain, the plot displays the genomes of GC cohort as circular chromosomes of 1,539,122 bp. (**B**) Circle diagram of the chromosome of *H. pylori* isolates strain, the plot displays the genomes of gastritis cohort as circular chromosomes of 1,591,513 bp. Circle (from outside to inside): circle 1 (virulence factor prediction); circle 2–7 (size of the genome in GC and gastritis cohort); circle 8 (size of the reference genomes); circle 9 (GC skew); circle 10 (GC content). GC, Gastric Cancer. GI, Gastritis. VF, Virulence Factor

significant differences in core SNPs between the GC and gastritis strains, while no substantial differences were observed in the homologous proteins between the two cohorts.

Synteny in genome organization of the 12 *H. pylori* isolates and the reference strains

To illustrate the utility of IslandCompare, we conducted an analysis focused on the prediction, comparison, and exploratory visualization of gene islands associated with GC, gastritis, and the reference of *H. pylori* strains. The gene-island prediction visualization features an interactive linear display covering each genome (Fig. 4). Gene islands are uniformly colored according to their sequence clusters, with the highlighted colors highlighting similar gene islands across all genomes. Through the results of gene island prediction, the location of each gene island, the relationship between each other and the possible evolution sequence of all strains in the cluster were visualized. The prediction results of *H. pylori* strains for GC and gastritis based on gene islands showed that the locations of gene islands for 91.7% (11/12) strains were

Table 2 Comparison of virulence factor prediction of *H. Pylori* from GC and gastritis

Class	Virulence factors	Related gene-GC	Related gene- Gastritis
Acid resistance	Urease	*UreA/E/F/G/H/I*	*UreA/B/E/F/G/H/I*
Adherence	AlpB (hopB)	*alpB/hopB*	-
	H. pylori adhesin A	*hpaA*	*hpaA*
	HopZ	*hopZ*	*hopZ*
	HorB	*horB*	*horB*
	Sialic acid binding adhesins	*sabA/hopP*	-
	Adherence-associated lipoprotein AlpA (hopC)	*alpA/hopC*	-
Immune evasion	Lipopolysaccharide Lewis antigens	*futB/C*	*futC*
Immune modulator	Neutrophil-activating protein (HP-NAP)	*napA*	*napA*
Motility	Flagella	*flaA/B/G, flagA/B/C/D/E1/E2/G2/H/I/K, flhA/B1/B2/F, fliA/D/E/F/G/H/I/L/M/N/P/Q/R/Y, pflA*	*flaA/B/G, flagA/B/C/D/E1/E2/G1/G2/H/I/K/L, flhA/B1/B2, fliA/E/F/G/H/I/L/M/N/P/Q/R/S/Y, motA/B, pflA*
Secretion system	Cag PAI type IV secretion system	*Cag1/3/4/5/C/D/E/F/G/H/I/L/M/N/P/Q/S/T/U/V/W/X/Z, virB11*	*Cag1/3/4/5/C/D/E/F/G/H/I/L/M/N/P/Q/S/T/U/V/W/X/Z, virB11*
	T4SS effectors cytotoxin-associated gene A	*cagA*	-
Toxin	Vacuolating cytotoxin	*vacA*	*vacA*

GC, Gastric Cancer

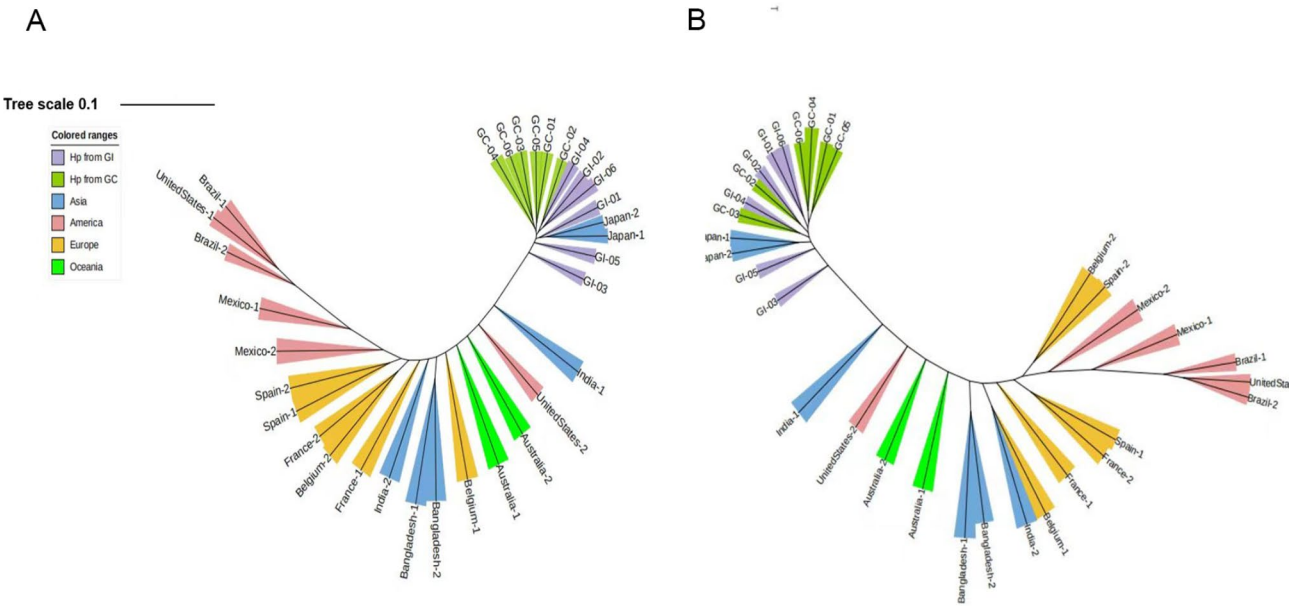

Fig. 3 Phylogenetic tree of the 12*H. pylori* isolates and the reference genomes. (**A**) Core SNP based phylogenetic tree of the 12 *H. pylori* isolates and the reference genomes. (**B**) Homologous protein based phylogenetic tree of the 12 *H. pylori* isolates and the reference genomes. GC, Gastric Cancer. GI, Gastritis. SNP, Single Nucleotide Polymorphism

concentrated in 1-1.4 Mb and were mainly divided into two distinct segments. In the upper segment, four strains (GC-05, GC-06, gastritis-02, and gastritis-05) cluster closely with gastritis-03 and India-1. Conversely, in the lower segment, five strains (GC-01, GC-02, gastritis-01, gastritis-04, and gastritis-06) are grouped with other strains of Asian origin, including Japan-2, India-2, and Bangladesh-1, as well as with GC-03 and GC-04. This clustering highlights the evolutionary relationships and shared genetic characteristics among the strains.

Genome homology and difference of *H. Pylori* in GC and gastritis patients

We used FastANI software to perform pairwise ANI analysis on 32 sorted *H. pylori* strains, and the results were visualized accordingly. In Fig. 5A, the ANI analysis indicates that the GC and gastritis strains examined in

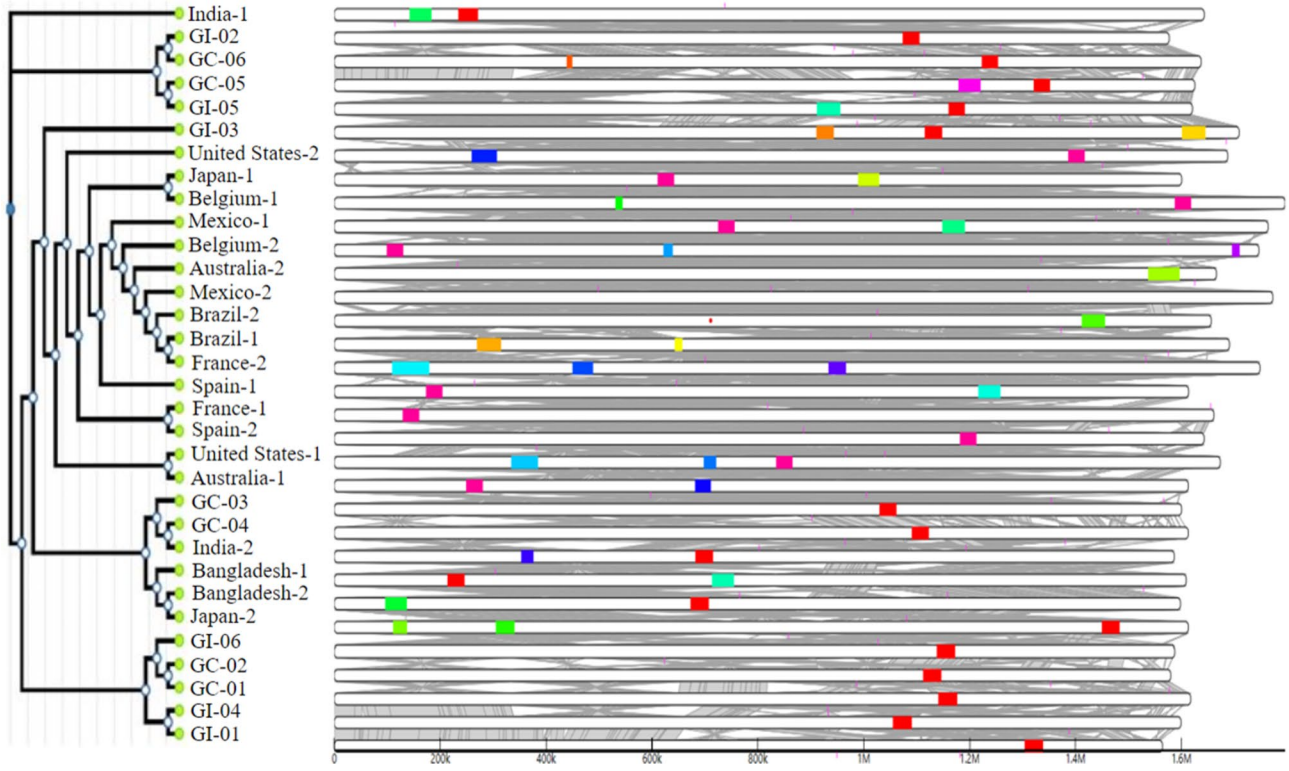

Fig. 4 Synteny in genome organization of the 12 *H. pylori* isolates and the reference strains. A phylogeny (left) indicates the relationship between the isolates in the analysis, with comparative visualization and zoom- in functionality available. GC, Gastric Cancer. GI, Gastritis

our study exhibit a high degree of homology with strains of Asian origin, including Japan-2, India-1, India-2, Bangladesh-1, and Bangladesh-2. Our analysis identified that the *H. pylori* strains associated with GC and gastritis share a total of 1,103 genes. Additionally, we found 160 genes that are specific to GC strains and 144 genes that are specific to gastritis strains (Fig. 5B).

The pan-genomic analysis of GC, gastritis, and reference strains revealed a substantial number of shared genetic components, with a total of 1,215 core genes identified across all strains. In contrast, the GC and gastritis strains exhibited a limited number of specific genes, ranging from 2 to 8 (Supplementary Figure S4A). GO functional annotation found that the specific genes associated with GC and gastritis were primarily concentrated in categories such as cellular anatomical entity, binding, cellular process, and metabolic process (Fig. 4C and Supplementary Figure S4B). Additionally, COG analysis found that the specific gene functions of GC and gastritis strains were mainly concentrated in metabolism, translation, biogenesis, replication, recombination, and repair (Supplementary Figure S4C-D). Furthermore, presented in Fig. 4D, indicated that the top four functions of specific genes of GC strains were nucleotide metabolism, amino acid metabolism, metabolism of cofactors and vitamins. For gastritis strains, the leading functions of their specific genes included the metabolism of cofactors and vitamins, translation, amino acid metabolism, and replication and repair (Supplementary Figure S4E).

Discussion

H. pylori exhibits a complex and long-term coexistence with humans, playing a significant role in the development of gastritis and GC [18, 19]. Researches have focused on whether the pathogenicity of *H. pylori* varies by population, source, or genetic specificity, and whether specific highly pathogenic strains exist [20, 21]. Therefore, identifying key factors contributing to strain pathogenicity and understanding region-specific highly pathogenic strains is essential. Here, we sequenced 12 *H. pylori* isolates of derived from GC or gastritis patients and performed whole genome–based comparative analysis to investigate the genetic structure and differences between GC and gastritis strains. Based on this, further in-depth research will continue, and it is hopeful that humanity will fully master the molecular biological characteristics of GC-causing strains, enabling targeted prevention and early diagnosis of GC, thereby revolutionizing strategies for the prevention and treatment of GC.

East Asia has long been a high-incidence region for gastritis and GC, with a strong link between *H. pylori* and their development [22, 23]. In this study, 12 *H. pylori* strains from gastritis and GC patients in China were selected as research objects, and it is representative to

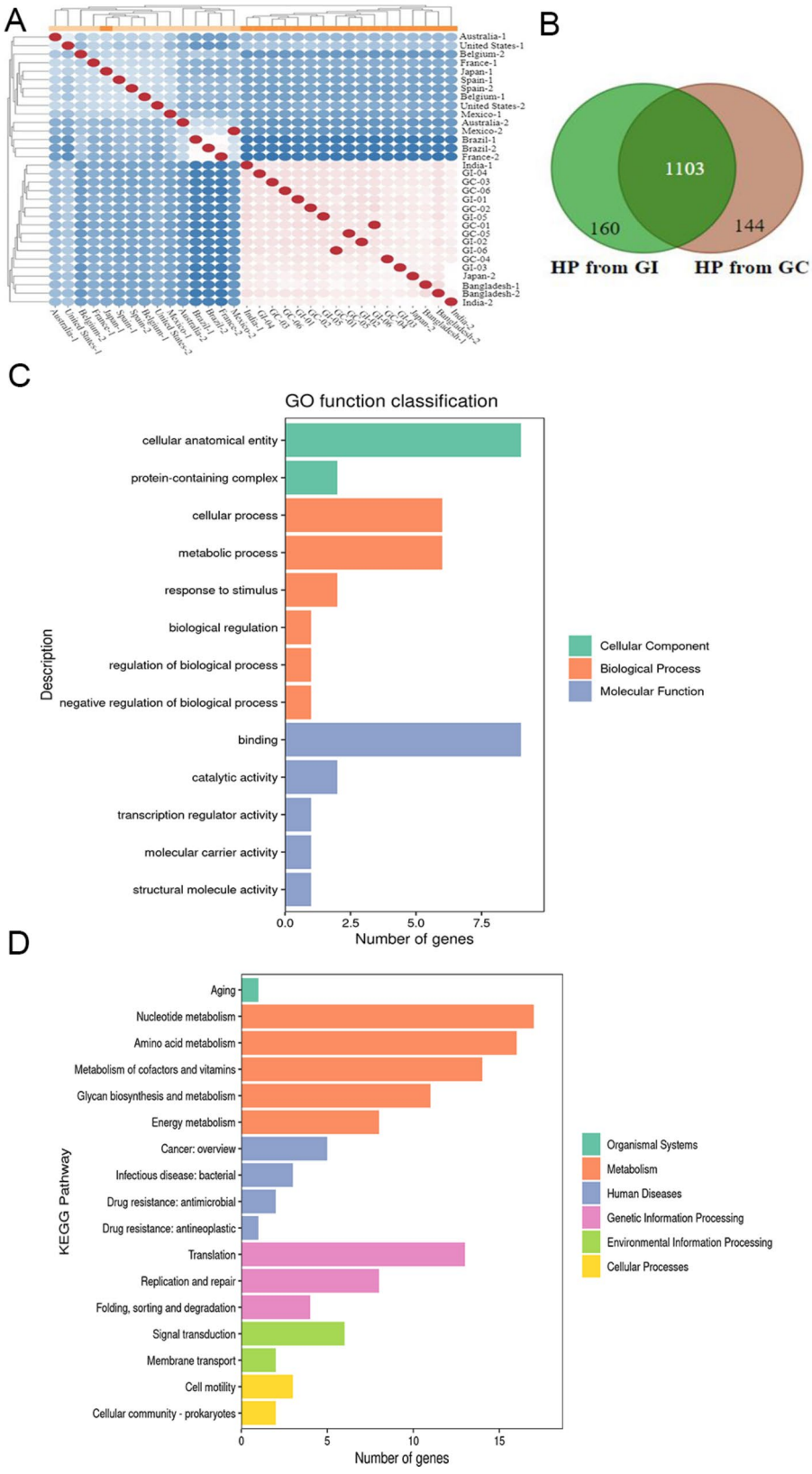

Fig. 5 Genome homology and difference of*H. pylori* in GC and gastritis patients. (**A**) Genome-wide ANI analysis of 12 *H. pylori* isolates and the reference strains. (**B**) Homologous and differential genes of *H. pylori* in patients with GC and gastritis. (**C**) GO analysis of *H. pylori* specific gene in GC patients. (**D**) KEGG analysis of *H. pylori* specific gene in GC patients. ANI, Average Nucleotide Homology. GO, Gene Ontology. KEGG, Kyoto Encyclopedia of Genes and Genomes. GC, Gastric Cancer. GI, Gastritis

explore the pathogenicity and genetic differences of the two cohorts. Meanwhile, we included representative strains from other regions worldwide sourced from public databases as references [18]. Consequently, the findings of our study will provide insights into the homology and difference between the strains of gastritis and GC, and with other reference strains to a certain extent.

In the present study, we used next-generation sequencing technology to sequence the whole genome of *H. pylori* strains that were isolated from gastritis and GC patients in Shanghai. The genome sequence of GC strain *H. pylori* consists of a ring chromosome with an average total length of 1,579,639 bp and contains approximately 1,594 CDS. In contrast, the gastritis strains exhibited a smaller average genome size of 1,569,508 bp and an average of 1,589 CDS, indicating lower genomic complexity compared to the GC group. Previous studies have reported that the genome size of *H. pylori* typically ranges from 1.5 to 1.9 Mb, with around 1,600 coding genes [24, 25]. Additionally, VFs of *H. pylori* from GC and gastritis, although with high consistency. However, the presence of the *cagA* gene sequence in GC strains, as compared to gastritis strains, may indicate a heightened level of virulence and aggressiveness in GC strains. While *cagA*-positive *H. pylori* strains are associated with various gastrointestinal conditions, including acute gastritis, peptic ulcer disease, and gastric cancer, our findings suggest that the presence of *cagA* in GC strains may reflect a specific propensity for increased virulence in the context of GC [26]. Globally, *cagA*-positive strains account for approximately 60% of *H. pylori* infections in individuals [27]. Phosphorylated *CagA* interacts with *SHP2*, *Csk*, *Crk* junction protein, and other proteins to activate the *ERK/MAPK/JNK/PI3K/JAK-STAT* signaling pathway, leading to abnormal expression of epithelial genes and inducing morphological changes in the "hummingbird phenotype" [28, 29].

In this study, the results of the SNP evolutionary tree revealed that the 12 *H. pylori* strains from patients with GC and gastritis in China exhibited variability in their lineages. The overlap between GC and gastritis *H. pylori* lineages sequenced in this study aligns with the co-evolution of *H. pylori* lineages observed in Japan, particularly among strains from gastritis patients. In addition, the phylogenetic tree based on homologous proteins, along with the visualization of gene island predictions, did not differentiate between *H. pylori* strains from GC and gastritis. The observation of two distinct lineages has been documented in previous studies. Previous genome-wide association studies of strains of hp-East Asia by Japanese scholars have shown that differences in SNPs between strains of GC and duodenal ulcer can be detected, and potential pathogenic mechanisms such as charge changes in ligand-binding pockets, changes in subunit interactions, and pattern switching DNA methylation have been proposed [30]. Yamaoka et al. identified virulence genes in the genomes of strains CHC155 (GC) and VN1291 (duodenal ulcer) as key risk factors for *H. pylori* pathogenicity [31]. Another study comparing *H. pylori* infections related to GC and duodenal ulcers suggested that *vacA* genotype status could help identify patients at high risk for developing GC [11]. The comparison results of strains VF described above in this study also found that the *cagA* status of *H. pylori* strains in GC and gastritis was different, which may be an important genomic feature of the two cohort strains. Understanding the implications of *cagA* status is crucial for early diagnosis and treatment of GC. The presence of specific *cagA* variants has been associated with increased virulence and a higher risk of progression from gastritis to GC [28]. By identifying patients with high-risk *cagA* variants, clinicians can implement targeted surveillance and preventive strategies, allowing for earlier intervention and potentially improving GC patient outcomes.

Subsequently, the results of pan-genomic and ANI analyses suggested that GC, gastritis and other reference *H. pylori* strains exhibited high levels of homology. This finding aligns with previous research, which has shown that *H. pylori* lineages are influenced by geographical regions and can be categorized into seven distinct lineages: HP-Africa1, HP-Africa2, HP-Sahul, HP-Europe, HP-Asia2, HP-Amerind and HP-East Asia [18]. However, it is anticipated that these coexisting lineages will converge over time due to the genomic plasticity of the strains [19]. In addition to genome homology, the gene function annotation results of COG, GO and KEGG further suggested that the *H. pylori* strains of GC and gastritis also had high similarity in gene function, and their specific gene functions primarily focused on processes related to metabolism, transcription, and repair. These findings are consistent with previous gene annotation studies conducted on strains from other regions [13, 32].

There are certain limitations to our study. First of all, due to the limited sample size in the study design and the regional nature of the samples, the obtained research results need to be further confirmed by large sample size and multi-center studies. Secondly, the reference genomes in the study only selected a few representative strains from each continent, which may also have a certain selection bias. Additionally, the selection of strains from GC and chronic atrophic gastritis patients presents another potential limitation. Given that chronic atrophic gastritis can be a precursor to gastric cancer, this choice may impact the accuracy of our findings. Indeed, further study including strains from patients with peptic ulcers, which represent another significant outcome of *H. pylori* infection, might provide a more comprehensive understanding of the pathogenicity of *H. pylori* across different

clinical conditions. Lastly, the differences between VF and SNP in strains of GC and gastritis have not been further explored, nor have relevant basic experiments been designed for verification, and more detailed analyses are needed in future studies to address these gaps.

Conclusion

In conclusion, GC and gastritis patient-derived *H. pylori* have some differences in VF and SNP, they also demonstrate a high degree of homology at other genomic levels within the Chinese population.

Abbreviations

vacA	Vacuolating cytotoxin A
cagA	Cytotoxin-associated gene A
SHP2	Src homology 2 domain-containing protein tyrosine phosphatase 2
Csk	C-src kinase
Crk	CT10 regulator of kinase
ERK	Extracellular signal-regulated kinase
MAPK	Mitogen-activated protein kinase
JNK	c-Jun N-terminal kinase
PI3K	Phosphoinositide 3-kinase
JAK	Janus kinase
STAT	Signal transducer and activator of transcription

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13493-6.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We thank Dr. He Teng for statistical advising and review of the manuscript.

Author contributions

Xu DZ and Kong PF designed the research study; Xu DZ, Kong PF, Yan YH, Duan YT, Fang YT, Dou Y and Xu YH performed the research; Kong PF, Yan YH, and Fang YT analyzed the data and wrote the manuscript; Xu DZ made the final review and revision of the manuscript; all authors have read and approve the final manuscript.

Funding

This work was supported by the Natural Science Foundation of China under grants 81972213.

Data availability

The whole genome sequence data of strains in this study have been deposited at SRA-NCBI. The NCBI BioProject accession number for WGS sequences reported in this paper is PRJNA1103397.

Declarations

Ethics approval and consent to participate

The study was reviewed and approved by Ethics Committee of Fudan University Shanghai Cancer Center Review Board [Approval No. FUSCC-D-2022-144]. All study participants or their legal guardian provided informed written consent about personal and medical data collection prior to study enrolment.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 February 2024 / Accepted: 10 January 2025
Published online: 07 April 2025

References

1. Li Y, Choi H, Leung K, Jiang F, Graham DY, Leung WK. Global prevalence of Helicobacter pylori infection between 1980 and 2022: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol. 2023;8(6):553–64.
2. Malfertheiner P, Camargo MC, El-Omar E. Helicobacter pylori Infect. 2023;9(1):19.
3. Smith SI, Suerbaum S, Usui Y. Helicobacter pylori, homologous-recombination genes, and gastric Cancer. Nat Reviews Disease Primers. 2023;388(13):1181–90.
4. Zhou XZ, Lyu NH. Large-scale, national, family-based epidemiological study on Helicobacter pylori infection in China: the time to change practice for related disease prevention. 2023;72(5):855–69.
5. Shirani M, Pakzad R, Haddadi MH, Akrami S, Asadi A, Kazemian H, Moradi M, Kaviar VH, Zomorodi AR, Khoshnood S, et al. The global prevalence of gastric cancer in Helicobacter Pylori-infected individuals: a systematic review and meta-analysis. BMC Infect Dis. 2023;23(1):543.
6. Bray F, Laversanne M, Sung H. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. 2024;74(3):229–63.
7. Kim JY, Kim N, Nam RH, Suh JH, Chang H, Lee JW, Kim YS, Kim JM, Choi JW, Park JG, et al. Association of polymorphisms in virulence factor of Helicobacter pylori and gastroduodenal diseases in South Korea. J Gastroenterol Hepatol. 2014;29(5):984–91.
8. You Y, He L, Zhang M, Fu J, Gu Y, Zhang B, Tao X, Zhang J. Comparative genomics of Helicobacter pylori strains of China associated with different clinical outcome. PLoS ONE. 2012;7(6):e38528.
9. Vaziri F, Najar Peerayeh S, Alebouyeh M, Mirzaei T, Yamaoka Y, Molaei M, Maghsoudi N, Zali MR. Diversity of Helicobacter pylori genotypes in Iranian patients with different gastroduodenal disorders. World J Gastroenterol. 2013;19(34):5685–92.
10. McClain MS, Shaffer CL, Israel DA, Peek RM Jr., Cover TL. Genome sequence analysis of Helicobacter pylori strains associated with gastric ulceration and gastric cancer. BMC Genomics. 2009;10:3.
11. El Khadir M, Alaoui Boukhris S, Benajah DA, El Rhazi K, Ibrahimi SA, El Abkari M, Harmouch T, Nejjari C, Mahmoud M, Benlemlih M, et al. VacA and CagA status as Biomarker of two Opposite End outcomes of Helicobacter pylori infection (gastric Cancer and duodenal ulcer) in a Moroccan Population. PLoS ONE. 2017;12(1):e0170616.
12. Romo-González C, Salama NR, Burgeño-Ferreira J, Ponce-Castañeda V, Lazcano-Ponce E, Camorlinga-Ponce M, Torres J. Differences in genome content among Helicobacter pylori isolates from patients with gastritis, duodenal ulcer, or gastric cancer reveal novel disease-associated genes. Infect Immun. 2009;77(5):2201–11.
13. Zeng X, Xiong L, Wang W, Zhao Y, Xie Y, Wang Q, Zhang Q, Li L, Jia C, Liao Y, et al. Whole-genome sequencing and comparative analysis of Helicobacter pylori GZ7 strain isolated from China. Folia Microbiol. 2022;67(6):923–34.
14. Lyu T, Cheung KS, Deng Z, Ni L, Chen C, Wu J, Leung WK, Seto WK. Whole genome sequencing reveals novel genetic mutations of Helicobacter pylori associating with resistance to clarithromycin and levofloxacin. Helicobacter. 2023;28(4):e12972.
15. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinf (Oxford England). 2015;31(4):587–9.
16. Yu Y, Zhang Z, Dong X, Yang R, Duan Z, Xiang Z, Li J, Li G, Yan F, Xue H, et al. Pangenomic analysis of Chinese gastric cancer. Nat Commun. 2022;13(1):5412.
17. Larson MA, Sayood K, Bartling AM, Meyer JR, Starr C, Baldwin J, Dempsey MP. Differentiation of Francisella tularensis subspecies and subtypes. J Clin Microbiol 2020;58(4).
18. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al. Traces of human migrations in Helicobacter pylori populations. Sci (New York NY). 2003;299(5612):1582–5.
19. Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Vigl EE, Malfertheiner P, Megraud F, et al. The 5300-year-old Helicobacter pylori genome of the Iceman. Sci (New York NY). 2016;351(6269):162–5.

20. Usui Y, Taniyama Y, Endo M, Koyanagi YN, Kasugai Y, Oze I, Ito H, Imoto I, Tanaka T, Tajika M, et al. Helicobacter pylori, homologous-recombination genes, and gastric Cancer. N Engl J Med. 2023;388(13):1181–90.
21. Shiota S, Suzuki R, Yamaoka Y. The significance of virulence factors in Helicobacter pylori. J Dig Dis. 2013;14(7):341–9.
22. Han Z, Liu J, Zhang W, Kong Q, Wan M, Lin M, Lin B, Ding Y, Duan M, Li Y, et al. Cardia and non-cardia gastric cancer risk associated with Helicobacter pylori in East Asia and the West: a systematic review, meta-analysis, and estimation of population attributable fraction. Helicobacter. 2023;28(2):e12950.
23. Russo AE, Strong VE. Gastric Cancer etiology and management in Asia and the West. Annu Rev Med. 2019;70:353–67.
24. Khosravi Y, Rehvathy V, Wee WY, Wang S, Baybayan P, Singh S, Ashby M, Ong J, Amoyo AA, Seow SW, et al. Comparing the genomes of Helicobacter pylori clinical strain UM032 and mice-adapted derivatives. Gut Pathogens. 2013;5:25.
25. Lamichhane B, Chua EG, Wise MJ, Laming C, Marshall BJ, Tay CY. The complete genome and methylome of Helicobacter pylori hpNEAfrica strain HP14039. Gut Pathogens. 2019;11:7.
26. Stein M, Rappuoli R, Covacci A. Tyrosine phosphorylation of the Helicobacter pylori CagA antigen after cag-driven host cell translocation. Proc Natl Acad Sci USA. 2000;97(3):1263–8.
27. Hatakeyama M, Higashi H. Helicobacter pylori CagA: a new paradigm for bacterial carcinogenesis. Cancer Sci. 2005;96(12):835–43.
28. Takahashi-Kanemitsu A, Knight CT. Molecular anatomy and pathogenic actions of Helicobacter pylori CagA that underpin gastric carcinogenesis. 2020;17(1):50–63.
29. Wang H, Zhao M, Shi F, Zheng S, Xiong L, Zheng L. A review of signal pathway induced by virulent protein CagA of Helicobacter pylori. Front Cell Infect Microbiol. 2023;13:1062803.
30. Tuan VP, Yahara K, Dung HDQ, Binh TT, Huu Tung P, Tri TD, Thuan NPM, Khien VV, Trang TTH, Phuc BH et al. Genome-wide association study of gastric cancer- and duodenal ulcer-derived Helicobacter pylori strains reveals discriminatory genetic variations and novel oncoprotein candidates. Microb Genomics 2021;7(11).
31. Phuc BH, Tuan VP, Binh TT, Tung PH, Tri TD, Dung HDQ, Thuan NPM, Fauzia KA, Tshibangu-Kabamba E, Alfaray RI, et al. Comparative genomics of two Vietnamese Helicobacter pylori strains, CHC155 from a non-cardia gastric cancer patient and VN1291 from a duodenal ulcer patient. Sci Rep. 2023;13(1):8869.
32. Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, Hanan F, Ramos R, Pereira U, Barh D et al. Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. BioMed research international. 2015:139580.

Publisher's note