Glycosaminoglycan biosynthesis pathway in host genome is associated with *Helicobacter pylori* infection

Dingxue Hu1,2,3, Yueqi Lu1,3, Daoming Wang1, Chao Nie1,2* & Yan Li1,2*

*Helicobacter pylori* is a causative pathogen of many gastric and extra-gastric diseases. It has infected about half of the global population. There were no genome-wide association studies (GWAS) for *H. pylori* infection conducted in Chinese population, who carried different and relatively homogenous strain of *H. pylori*. In this work, we performed SNP (single nucleotide polymorphism)-based, gene-based and pathway-based genome-wide association analyses to investigate the genetic basis of host susceptibility to *H. pylori* infection in 480 Chinese individuals. We also profiled the composition and function of the gut microbiota between *H. pylori* infection cases and controls. We found several genes and pathways associated with *H. pylori* infection (*P* < 0.05), replicated one previously reported SNP rs10004195 in TLR1 gene region (*P* = 0.02). We also found that glycosaminoglycan biosynthesis related pathway was associated with both onset and progression of *H. pylori* infection. In the gut microbiome association study, we identified 2 species, 3 genera and several pathways had differential abundance between *H. pylori* infected cases and controls. This paper is the first GWAS for *H. pylori* infection in Chinese population, and we combined the genetic and microbial data to comprehensively discuss the basis of host susceptibility to *H. pylori* infection.

More than half of the world population were infected by *Helicobacter pylori*1. It is the major cause of gastritis (80%) and gastroduodenal ulcer disease (15–20%) and the only bacterial pathogen believed to cause gastric cancer2,3. The prevalence of *H. pylori* infection has large amounts of variations among countries. One of the causes may be that *H. pylori* populations are very diverse, due to point mutations and larger substitutions, insertions or deletions that may involve one or more genes or multigene segments (including restriction-modification genes and at least one pathogenicity island). Divergence among *H. pylori* strains would be further increased by differences among people in traits important to individual strains4. Numerous epidemiological studies have investigated the associations between polymorphic genetic elements of *H. pylori* and gastric disease development5,6. One study found that gastric cancer was associated with the presence of babA and genes in the *cag* pathogenicity island of *H. pylori* through comparing the strains isolated from gastric cancer patients and gastritis individuals5. In addition to the variation of *H. pylori* genome, the host genetic variation also contributed to the *H. pylori* susceptibility7,8, the reported heritability was 57%8. So far, only one genome-wide association study (GWAS) of *H. pylori* infection was performed in European population9. Genetic variations in the toll-like receptor (TLR) locus were found to be associated with *H. pylori* seroprevalence. TLRs are important pattern recognition receptors on cellular surfaces in the innate immune system10, known to be essential in providing protective immunity against infection. Many studies reported that TLRs were associated with *H. pylori* infection11,12. On the *H. pylori* side, the immune response is triggered by bacterial membrane components, including lipopolysaccharides and lipid A, as well as cytotoxins and *H. pylori* urease activity13. Several genetic studies have demonstrated that the variants in *IL1B* were associated with susceptibility to gastric cancer among *H. pylori* infected individuals, this suggested that immune-related genes may be involved in *H. pylori* infection14,15. However, there is no GWAS for *H. pylori* infection on Chinese population.

On the other hand, gut microbiota was reported to play a role in regulating host metabolic function, protecting against pathogen infections and promoting immune system maturation. Previous studies reported some alterations of gut microbiota in *H. pylori* infected patients16,17. But there is a lack of research putting genetic...
susceptibilities and gut microbiome study together to integrate more information for understanding the mechanism of *H. pylori* infection.

Therefore, we carried out a GWAS on *H. pylori* infection and also compared the profiles of gut microbiota between cases and controls with the aim of identifying genetic variants influencing *H. pylori* infection and gaining more insights into the mechanism of *H. pylori* in Chinese population.

## Results

### Study sample.

224 *H. pylori* infected positive individuals and 256 negative individuals were selected from an existing data set with the results of carbon-13 urea breath test (C13 test) for *H. pylori* infection. The written form of consent was signed for each individual. The C13 test results include diagnosis of *H. pylori* infection and the delta over baseline (DOB) values for cases. The distribution of DOB values for cases was shown in Supplementary Fig. S1. The detailed characteristic statistics of the samples were described in Table 1. The case and control groups were designed to match age, sex and body mass index (BMI). There were no significant differences between two groups in age, sex and BMI.

### Association analysis on *H. pylori* infection status and DOB measurements.

In order to identify genetic loci involved in *H. pylori* infection, we constructed two models in GWAS after standard quality control (QC) processes. Logistic regression was performed in 224 cases and 256 controls and linear regression was performed in 224 cases. The scree plot of principal component analysis (PCA) was shown in Supplementary Fig. S2a. The top 2 components could explain most of the variations of the population and no clear population stratification was observed (Supplementary Fig. S2b). The variant-based associations showed no genome-wide significant single nucleotide polymorphisms (SNPs) (*P* < 5e−8). There were still SNPs reached suggestive level (*P* < 1e−5) (Fig. 1), 10 SNPs for logistic regression and 59 SNPs for linear regression (Supplementary Table S1 and Supplementary Table S2). The Q–Q plots of both associations were shown in Supplementary Fig. S3. The inflation factor was 1.00445 for case/control analysis and 1.01076 for DOB values association, indicating there were no population stratification for the analyses.

Then we further investigated the regions harbor suggestive significant SNPs. In case/control association, two signals were located in genes LIN20211 (Fig. 2a) and CACNA1B (Fig. 2b) regions, another signal was located in intergenic region between *NEDD1* and *RMST* (Supplementary Table S1 and Supplementary Fig. S4). Copy number variation in *LIN20211* region has been reported to be associated with autism spectrum disorder26,27. This region also harbored another non-coding gene *PGD5-AS1*, it was reported to regulate gastric cancer cell proliferation and chemoresistance28. Protein encoded by gene *CACNA1B* was involved in calcium channel activity. This gene has been found recurrently mutated in 2 or more tumor tissues of 5 gastric cancer patients with *H. pylori* infection29. In linear regression based GWAS, 59 SNPs passed the nominal significance level (*P* < 1e−5). The signals were located in genes *SLC25A3P1, DMRTB1, ST6GALNAC3, KIDINS220, DPP10, CELF2, CCDC77, FTO, ATP2C2-AS1, LARGE1* regions and some intergenic regions. Their summary statistics were shown in Supplementary Table S2 and Supplementary Fig. S5. Among these regions, *FTO* (Fig. 2c) was strongly associated with obesity and BMI22. Obesity has been reported to be associated with *H. pylori* infection23. It is also a risk factor for gastric cancer24. One study found that gastric cancer tissues had high *FTO* expression as compare with the adjacent non-cancerous tissues25. LARGE1 gene (Fig. 2d) was involved in immune response for disease pathogens26.

### Set-based association for further identification of susceptibility genes and pathways for *H. pylori* infection.

To improve GWAS power, set-based analyses were performed, which combined the GWAS summary statistics of all variants within a putative biologically functional unit (coding region and possible regulatory region, genes within one pathway) to obtain a single *P* value that represented the significance level of the associations between *H. pylori* infection and the functional unit. Our SNPs were mapped onto 25,928 genes for gene-based association. The top associated genes for case/control comparison were *LOC105376331* and *LIN20211* (gene-level *P* < 1e−04). And the most significant genes for linear regression were *LARGE1* (gene-level *P* < 1e−04). They failed to pass genome-wide gene-based significance threshold (0.05/25928). Then we further mapped these genes to canonical pathways from molecular signature database27,28. 164 pathways were included in the analyses. Top pathways associated with *H. pylori* infection (*P* < 0.05) were listed in Tables 2 and 3. In the logistic model-based pathway association, the glycosphingolipid biosynthesis related pathway (glycosphingolipid biosynthesis—lacto and neolacto series) was the most significant pathway.

|       | Number (N) | Sex female count (percent) | Male count (percent) | Age Mean (SD) | Range | BMI Mean (SD) | Range |
|-------|------------|---------------------------|----------------------|---------------|-------|---------------|-------|
| Case  | 224        | 127 (56.70%)              | 97 (43.30%)          | 30.78 (5.46)  | 22–58 | 22.01 (3.00)  | 15.20–30.60 |
| Control | 256        | 136 (53.13%)              | 120 (46.88%)         | 31.18 (5.40)  | 22–69 | 22.25 (3.22)  | 15.70–37.80 |
| Total | 480        | 263 (54.79%)              | 217 (45.21%)         | 30.99 (5.46)  | 22–69 | 22.14 (3.12)  | 15.20–37.80 |

Table 1. Characteristic statistics of the study population. After Wilcoxon test, there was no significance in age difference between case and control. In the logistic regression model, *H. pylori* infection status ~ sex, we found there was no significant correlation between sex and *H. pylori* infection. BMI had no significant difference in the cases and controls.
One study reported that after \textit{H. pylori} eradicated successfully, glycosphingolipid biosynthesis—lacto and neolacto series pathway was significantly altered\textsuperscript{29}. In the linear model-based pathway association, the most significant pathway was antigen processing and presentation pathway. Glycosaminoglycan biosynthesis of heparan sulfate (HS) pathway was significant in the both two models, it was a pathway correlated with bacterial adherence\textsuperscript{30}. Several immune-related pathways were ranked top in the pathway-based association, indicating the variation of immune systems leading to different infection status of \textit{H. pylori} infection. Folate biosynthesis pathway was also significantly correlated with different infection status of \textit{H. pylori} infection. Study reported that folate level was significantly decreased (18.6\%) in \textit{H. pylori} infected individuals than those who were not infected\textsuperscript{31}.

\textbf{Replication of previous reported \textit{H. pylori} infection related genes.} We have checked the two signals reported in previous GWAS in European population\textsuperscript{9}. The rs10004195 in gene \textit{TLR1} was replicated in
Figure 2. Regional association plots for the regions containing associated genes. Regional association plots for the LINC02011 (a), CACNA1B (b), FTO (c) and LARGE1 (d). LINC02011 (a) and CACNA1B (b) were identified based on the GWAS results from the H. pylori infection cases versus controls, FTO (c) and LARGE1 (d) were identified based on the DOB values association study of H. pylori infection cases. The color of the variants is based on the linkage disequilibrium with rs1687312 (LINC02011) (a), rs7875160 (CACNA1B) (b), rs8046411 (FTO) (c) and rs5749679 (LARGE1) (d). For all plots, each point represents a SNP, where the x axis represents the position of the SNPs and the y axis represents the $-\log_{10} P$ values of the GWAS results. Each point is color-coded with the r$^2$ value as calculated with the source of LD information retrieved from hg38/1000 Genomes Nov 2014 EAS (Asian). Plots showed the most significant SNPs flank 300 kb.

Table 2. Candidate pathways associated with H. pylori infection in logistic regression model. No.SNPs: number of SNPs in the pathway; Chisq(Obs): sum of chi-squared test-statistics of all SNPs in the set; $P$ value: pathway set-based test $P$ value; TopSNP.Pvalue: smallest single-SNP GWAS $P$ value in the pathway set; TopSNP: the top associated GWAS SNP. The rsID is based on dbSNP build 138.
sugar diet, and it was closely related to unhealthy fasting serum lipid profile in obese women. This species categories from kingdom to species assigned by MetaPhlAn2. Among them, 782 categories were classified to infected cases and 218 controls. In total, we identified 961 taxonomic differences of gut microbiomes were identified between \( H. pylori \) infected cases and \( H. pylori \) infected negative; HPI positive: \( H. pylori \) infection in Wilcoxon test. Taxo: identified bacteria names at last taxonomic level by MetaPhlAn2; Mean difference: difference of bacteria abundance between \( H. pylori \) infection positive. The SNP rs12204338 was located in a region which included multiple TLRs (TLR1, TLR2, and TLR6). TLRs play crucial roles in activating innate immunity. The SNP rs368433 was located in \( FCGR2A \) gene region. Further gene-based association combining the information from neighboring SNPs within a region showed that TLRs region was replicated (\( P < 0.05 \)) based on the linear regression results. \( FCGR2A \) was not replicated (\( P > 0.05 \)).

### Differences of gut microbiomes were identified between \( H. pylori \) infected groups and controls.

Stool samples from 403 individuals were collected and metagenome sequencing was utilized to study the gut microbiome between 185 \( H. pylori \) infected cases and 218 controls. In total, we identified 961 taxonomic categories from kingdom to species assigned by MetaPhlAn2. Among them, 782 categories were classified to different taxonomic levels. After removing categories with carriers less than 10%, 301 categories were remained.

| Pathway set | No.SNPs | Chisq(Obs) | P value | TopSNP.Pvalue | TopSNP |
|-------------|---------|------------|---------|---------------|--------|
| KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION | 46 | 126.12 | 9.74E-03 | 1.03E-03 | rs12204338 |
| KEGG_AUTOIMMUNE_THYROID_DISEASE | 46 | 126.12 | 9.74E-03 | 1.03E-03 | rs12204338 |
| KEGG_HEDGHEOG_SIGNALING_PATHWAY | 6 | 23.2668 | 1.60E-02 | 5.35E-03 | rs9576171 |
| KEGG_TEREPENOID_Backbone_BIOSYNTHESIS | 9 | 32.6463 | 1.91E-02 | 8.99E-03 | rs16836189 |
| KEGG_NEOACTIVE_LIGAND_RECEPTOR_INTERACTION | 29 | 70.052 | 2.28E-02 | 1.36E-03 | rs2146490 |
| KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY | 19 | 48.4956 | 2.97E-02 | 2.09E-02 | rs12602449 |
| KEGG_JAK_STAT_SIGNALING_PATHWAY | 19 | 48.4956 | 2.97E-02 | 2.09E-02 | rs12602449 |
| KEGG_FOLATE_BIOSYNTHESIS | 21 | 46.2834 | 4.66E-02 | 1.16E-02 | rs13131273 |
| KEGG_GLUCOSAMINOLYCAN_BIOSYNTHESIS.HEPARAN_SULFATE | 80 | 135.496 | 4.77E-02 | 5.04E-03 | rs12195272 |

Table 3. Candidate pathways associated with \( H. pylori \) infection in linear regression model. No.SNPs: number of SNPs in the pathway; Chisq(Obs): sum of chi-squared test-statistics of all SNPs in the set; P value: pathway set-based test P value; TopSNP.Pvalue: smallest single-SNP GWAS P value in the pathway set; TopSNP: the top associated GWAS SNP. The rsID is based on dbSNP build 138.

| Taxo | Mean difference | P value | FDR-adjusted P | Presence ratio |
|------|-----------------|---------|----------------|---------------|
| Holdemanella biformis | 0.0008 | 0.0004 | 0.0579 | 0.2156 | 0.3838 |
| Catenibacterium mitsuokai | 0.0002 | 0.0010 | 0.0785 | 0.0596 | 0.1622 |
| ‘Erysipelotrichaceae noname’ | 0.0008 | 0.0003 | 0.0212 | 0.4266 | 0.5946 |
| Catenibacterium | 0.0002 | 0.0010 | 0.0349 | 0.0596 | 0.1622 |
| Prevotella | 0.0648 | 0.0037 | 0.0861 | 0.8716 | 0.8973 |

Table 4. Bacteria associated with \( H. pylori \) infection in Wilcoxon test. Taxo: identified bacteria names at last taxonomic level by MetaPhlAn2; Mean difference: difference of bacteria abundance between \( H. pylori \) infection positive group and \( H. pylori \) infection negative group; P value: P value of Wilcoxon test; FDR: taxonomic level-based FDR adjusted P value; Presence ratio: ratio of non-zero abundance individuals to all people, HPI negative: \( H. pylori \) infection negative; HPI positive: \( H. pylori \) infection positive.
Furthermore, we identified 37 *H. pylori* infection specific metabolic pathways which have different abundance in the case and control groups through logistic regression at nominal significance level (P < 0.05) (Supplementary Table S4). These results suggested that pathways involved in amino acids biosynthesis (e.g.: L-isoleucine, L-ornithine), were observed in *H. pylori* infection negative individuals, while L-lysine, L-lysine biosynthesis, S-adenosyl-L-methionine cycle I, L-citrulline metabolism and L-histidine degradation pathways had higher abundance in *H. pylori* infection positive groups), nucleotides metabolism (e.g.: pyrimidine related pathways were enriched in *H. pylori* infection negative group, while purine related pathways were enriched in *H. pylori* infection positive groups.), and other energy metabolites (e.g.: pyridoxal 5'-phosphate biosynthesis, glycolysis, homolactic fermentation, mannitol cycle, folate related pathways and peptideglycan pathways) may be different between *H. pylori* infection case and control groups. Interestingly, we found that folate biosynthesis pathway was associated with *H. pylori* infection in both our genetic and microbiome study. It has been reported that folate level was significantly decreased (18.6%) in *H. pylori* infection individuals than those who were not infected in a low-income population in the United States.

**Discussion**

In this study, we performed genome-wide association study on 224 *H. pylori* infected individuals and 256 negative individuals and also compared their differences of gut microbiome with the aim of gaining more insights into the bases and consequences of *H. pylori* infection. We identified 3 genetic regions which were associated with *H. pylori* infection status and 14 genetic regions which were related to the bacteria load at suggestive significance level (P < 1e−5). Further pathway-based analysis showed that several pathways were associated with *H. pylori* infection. Those pathways were related to immune response and glycosylation. One of the two reported loci from previous GWAS analysis was replicated in our data. In gut microbiome analysis, 2 species, 3 genera and several pathways were identified to have different abundance between *H. pylori* infected cases and controls.

While comparing the results with the reported GWAS about *H. pylori* infection in 2013, we found that rs10004195 in gene TLR1 was associated with the severity of *H. pylori* infection (P = 0.002023), but it was not associated with the infection status. TLRs are known to be essential in providing protective immunity against infection. The minor allele of rs10004195 may be associated with less efficiency of anti-inflammatory TLR1–TLR2 signaling. The other reported SNP rs368433 was filtered out by QC in our genetic data due to its low frequency in Chinese population (MAF = 0.009).

One possible cause of these different results may be the distinct methods for identifying *H. pylori* infection. In our study, we adopted C13 test to detect *H. pylori* infection. While in the GWAS for European population, they tested anti-*H. pylori* serum IgG antibody titers. C13 test has higher sensitivity and specificity than anti-*H. pylori* serum IgG antibody titers test. Although, anti-*H. pylori* serum IgG antibody titers could represent previous infection, it can hardly represent the current bacteria load in an individual. Besides, C13 test is a more convenient handling method. Therefore, we have chosen C13 breath tests in our study to measure both onset and severity of *H. pylori* infection. In addition, the population ancestry differences could be another cause for the variation of GWAS results. SNP rs368433 is common in European population and significantly associated with *H. pylori* infection. However, it is a low frequency SNP (MAF = 0.009) in Chinese population, and it was filtered out in our QC processes. We also checked the adjacent SNPs whose position are close to rs368433, they were not associated with *H. pylori* infection in our data. The true causal variant could still be hidden in this genomic region, but due to different linkage disequilibrium (LD) structure among European and Chinese populations, the alleles tagging true causal variants may be different. Future fine mapping will be required to extend the region and dig the hidden information.

Another difference for *H. pylori* infection between Chinese and European population is that they carry distinct *H. pylori* strains. Almost all Asian *H. pylori* strains contain the complete cog-pathogenicity island (PAI) which was associated with more severe disease in *H. pylori* infected individuals.

When comparing our results to existing functional studies, the functions of our newly identified genes and pathways were highly relevant to *H. pylori* infection. FGDS-AS1 regulates cancer gastric cell proliferation. CAC-NA1B had accumulated somatic mutations in gastric tumor with *H. pylori* infection. Genetic variants associated with obesity were also ranked top in our results. FTO is one of the leading genes associated with obesity and BMI. Studies also found that gastric cancer tissues had high FTO expression compared with the adjacent non-cancerous tissues. Obesity was considered as a risk factor for *H. pylori* infection. Interestingly, differences of obesity related bacteria were also identified in our gut microbiome analysis. *C. mitsuokai* was enriched in *H. pylori* infection positive group, which was also enriched in obese women. The associations among obesity, genetics of obesity, obesity related bacteria and *H. pylori* infection were reported pair wisely, but their causal relationships are still largely unknown. It is reported that *H. pylori* persistent infection had a negative effect on the fall of BMI values in Chinese obese population. This comorbidity between *H. pylori* infection and obesity may be caused by the pleiotropic effects of FTO and other genes. We also found that glycometabolism-related genes and pathways were associated with *H. pylori* infection. LARGE1 encoded a glycosylase, which played an important role in host–pathogen interactions. Our pathway-based associations identified that glycosphingolipid biosynthesis and glycosaminoglycan biosynthesis were associated with *H. pylori* infection. Glycosphingolipids (GSLs) are components of cell-surface or basement membrane-associated proteoglycans produced by both epithelial and mesenchymal cells. They are known to play a role as receptors in maulgen invasion. Glycosaminoglycans have been shown to be recognized by *H. pylori* outer-membrane proteins and to be involved in the adhesion of the bacteria in cell-line models. In the functional pathway enrichment of microbiome association study, we found that glycosylation related pathways were correlated with *H. pylori* infection. Many studies found that *H. pylori* can metabolize glucose by both oxidative and fermentative pathways. *H. pylori* infected gastric epithelial cells have exhibited increased glycosylation and increased expression of Lon protease 1 (Lonp1), Lonp1 is a key
regulator of metabolic reprogramming in *H. pylori*-induced gastric carcinogenesis\(^4,5,55\). We also identified two immune-related pathways, including antigen-processing and presentation pathway as well as autoimmune thyroid disease pathway. Study has reported that *H. pylori* affects the antigen presentation activity through microRNA\(^56\). *H. pylori* infection also has immune evasion strategies. Therefore, it was as expected that the genetic variations of immune-related genes may affect the susceptibilities of *H. pylori* infection.

While extensive number of studies have been focused on the alterations of gastric microbiota\(^29\) since *H. pylori* mainly colonizes the stomach\(^37\), little is known about the impact of *H. pylori* on gut microbiota. We have performed association analyses between *H. pylori* infection status and gut microbiome using shotgun metagenome sequencing. Compared to existing 16S rRNA gene sequencing for gut microbiota, shotgun sequencing has more power to identify less abundant taxa than 16S sequencing\(^26\). We identified 2 species, 3 genera and several pathways had differential abundance between *H. pylori* infected cases and controls. Interestingly, we found that folate biosynthesis pathway was associated with *H. pylori* infection in both genetic and microbiome studies. Reduction of vitamin B12 and folate concentrations has been reported in *H. pylori*-infected patients\(^39\). Folate acid was synthesized by bacteria\(^60\). Altogether, it is suggested that *H. pylori* infection could alter gut microbiome and thereby influence the nutrition supply of the host. More metabolomic and nutrition studies may be required to study the consequences of *H. pylori* infection.

There are some limitations of current study. The sample size was not large enough to boost the signals to genome-wide significance. We have leveraged on the set-based association by combining the SNPs within a biological meaningful set and generated a summed statistic for the whole set. Our set-based association identified glycosaminoglycan and immune-related pathways, which were functional relevant with *H. pylori* infection. We also studied the gut microbiome differences between *H. pylori* infected cases and controls, the results showed more supporting evidence that the glycosylation related function was involved in *H. pylori* infection. Together, we provided a list of candidate genes, pathways and bacteria associated with *H. pylori* infection for further replication and validation.

**Methods**

**Sample recruitment and data collection.** The samples of current study were selected from a cohort study of Shenzhen local area. Individuals who had C13 test and no symptoms of any gastrointestinal conditions were selected. C13 test is used for detecting *H. pylori* infection. It involves taking C13 isotope-labeled urea molecules by mouth, and then measuring the changes in carbon dioxide in the exhaled carbon by an isotope mass spectrometer. The main measurement index was the δ value, which measured the \(^{13}C/^{12}C\) ratio in the breath samples. DOB value is the difference between post-drug δ value minus pre-drug δ value. *H. pylori* infection is considered positive if DOB ≥ 4. Otherwise, *H. pylori* infection is negative (DOB < 4). Individuals who were diagnosed as *H. pylori* infection for both year 2017 and 2018 were further considered as the cases in our study. Individuals with DOB value equal to 0 were defined as control to ensure negative infection. At the end, 480 subjects (female N = 263, male N = 217) were selected for the association analyses.

**Whole genome sequencing.** The whole blood drawn from the participant vein was stored in the ethylenediaminetetraacetic acid (EDTA) anticoagulant tubes to avoid hemolysis, while the plasma was obtained by centrifugation (3000 rpm, 10 min) and was preserved at \(\sim 8^\circ C\) until assay. The white cells were isolated for genomic DNA extraction. Whole-genome sequencing (WGS) to 30x were conducted on BGI-seq500 sequencer. WGS data were aligned and variants called by the Picard/BWA\(^44/\)GATK\(^62\) pipeline. SNPs with mapping quality greater than 40, sequencing depth greater than 4, variant quality greater than 2.0, Phred score of Fisher's test \(P\) value for stand bias smaller than 60.0, haplotype score smaller than 13.0 and distance of alternative allele from the end of reads greater than 8.0 were kept for following analyses.

**GWAS quality control.** The common practice of GWAS QC is conducted through two dimensions of the data: samples and variants. All QC assessments and successive filtering were done using PLINK1.9\(^63\). Samples had more than 1% missing genotyped SNPs, different genetic sex with the record in the phenotypic database, abnormal autosomal heterozygosity were removed. Sex conflict and abnormal heterozygosity may be caused by DNA sample contamination. One of paired samples who have genetic relationship within three degree of relatedness was filtered out. SNPs had high rate of missing genotypes, and deviated from the Hardy–Weinberg equilibrium (HWE) test \(P \leq 1\times 10^{-05}\) as well as SNPs on X and Y chromosomes and mitochondria were also removed. After variant QC, 5,169,391 common SNPs (MAF > 0.05) were including for the association analyses. PCA was performed using SNPs on autosomal chromosome by PLINK1.9\(^63\) to investigate population stratification. No clear sub-cluster was observed. Typical north to south Grandunt was demonstrated by the first principal component (PC). Linear and logistic regression added top two PCs as covariates.

**Single-variant-based association.** When phenotypes and genotypes have passed all the QC processes, the statistical analysis of disease associations can be performed. The basic analysis of genome-wide association data is a series of single-locus statistical tests, examining each SNP independently for association to the disease status\(^64\). The types of statistical tests utilized mostly depend on the types of disease phenotypes. For continuous phenotype, DOB values of the C13 test results, additive model into linear regression was performed in 224 cases to identify genes associated with severity of *H. pylori* infection. While for binary trait, the case/control status, additive model into logistic regression was utilized in 224 cases and 256 controls to search for genes related to susceptibility of *H. pylori*. According to the scree plot of PCA (Supplementary Fig. S2), the top 2 PCs could explain most of the variations of the population and no clear population stratification observed (Supplementary Fig. S2). After Wilcoxon test, there was no significant difference for age in case/control and cases. In the logistic
regression model, *H. pylori* infection status ~ sex, we found there was no significant correlation between sex and *H. pylori* infection. Therefore, the association analyses were only adjusted the top 2 PCs. The regression model used in DOB values analysis was DOB values ~ Genotype + PC1 + PC2. The regression model used in *H. pylori* infection was *H. pylori* status ~ Genotype + PC1 + PC2. The adaptive permutation tests were used to generate significance levels empirically. All the analyses were performed in PLINK1.9.1 The Manhattan plots were generated in R 3.6.1. Regional plots were generated by LocusZoom tool (http://locuszoom.org/).

**Gene/pathway-based association.** The standard analysis of genome-wide association study uses single SNP marker as the test unit. However, due to small sample size, the complex LD structures and ethnic differences among different populations, many replication studies have failed. To improve GWAS power, the gene-based association has been proposed.66 The gene-based analysis combined the summary statistics of all variants within a putative gene (coding region and possible regulatory region) to obtain a single P value that represents the significance level of diseases and genes. Similarly, variants within genes on the same pathway were combined for pathway-based analysis. Gene-based and pathway-based analyses were performed using GCTA-fastBAT tool.67,68 In gene-based analysis, a gene region was defined as ± 8 kb of UTRs of a gene.

**Metagenomics sequencing.** Fresh stool samples were collected from recruited volunteers for metagenomics sequencing. Theecal DNA extractions were processed following the MetaHIT protocol, then single-end metagenomics sequencing was performed using BGISEQ-500 platform. The low-quality reads were discarded, and the host DNA were removed based on human genome reference (hg38) by SOAP269 (version 2.22). Taxonomic analysis was performed using MetaPhlAn225 following removal of human reads. The relative abundance of species and genera were used in the current study. Meanwhile, gene family abundance and pathway abundance of each sample were profiled using HUMAnN270 with default parameters. HUMAnN270 reported the abundance of gene families from the UniProt Reference Clusters (UniRef90), which were further mapped to microbial pathways from the MetaCyc metabolic pathway database. Gut microorganisms belonging to Kingdom Bacteria were extracted for correlation analysis between *H. pylori* infection and microbial abundance. Bacteria were excluded if they have less than 10% carriers in the population. After filtering, the abundance of each bacterium was rescaled into percentage within an individual. Then, Wilcoxon test was used to compare the abundance of bacteria between *H. pylori* cases and controls. The P values at the same taxonomic levels were adjusted by FDR. Alpha diversity assessed the bacteria diversity within individuals was compared in different *H. pylori* infection status groups using Wilcoxon test. Beta diversity illustrating distance between individuals in terms of bacterial composition was demonstrated as principal coordinates analysis (PCoA) plot. Alpha and beta diversity analyses were conducted in R 3.6.2 with vegan,71 phyloseq,72 and microbiome packages.73 Additionally, we dichotomized the abundance data to represent whether a bacterium is present in an individual. Logistic regression adjusted for age and sex was then used to analyze the associations between the presence of bacteria and the status of *H. pylori* infection. FDR method was used for multiple testing adjustment. In pathway analysis, pathways present in over 10% of samples were selected for subset analyses.

**Ethnics statement.** The study protocol was reviewed and approved by the Institutional Review Board of BGI (No. BGI-IRB 16068). Review procedures in BGI-IRB meet GCP principles and relevant national regulations for this application. Written informed consent was obtained from all the study participants prior to their enrolment.

**Data availability**

The data supported the findings of this study has been deposited into CNGB Sequence Archive (CNGB) of China National GeneBank DataBase (CNGDb) with accession number CNP0001557.

Received: 26 May 2021; Accepted: 31 August 2021
Published online: 14 September 2021

**References**

1. Hooi, J. K. et al. Global prevalence of Helicobacter pylori infection: Systematic review and meta-analysis. Gastroenterology 153, 420–429 (2017).
2. Montecucco, C. & Rappuoli, R. Living dangerously: How Helicobacter pylori survives in the human stomach. Nat. Rev. Mol. Cell Biol. 2, 457–466 (2001).
3. Kusters, J. G., Van Vliet, A. H. & Kuipers, E. J. Pathogenesis of *Helicobacter pylori* infection. Clin. Microbiol. Rev. 19, 449–490 (2006).
4. Suerbaum, S. & Josenhans, C. Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat. Rev. Microbiol. 5, 441–452 (2007).
5. Berthelot, E. et al. A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. BMC Biol. 16, 1–11 (2018).
6. Whitmire, J. M. & Merrell, D. S. Helicobacter pylori genetic polymorphisms in gastric disease development. Adv. Exp. Med. Biol. 1149, 173–194 (2019).
7. Graham, D. Y. et al. Epidemiology of Helicobacter pylori in an asymptomatic population in the United States: Effect of age, race, and socioeconomic status. Gastroenterology 100, 1495–1501 (1991).
8. Malaty, H. M., Enstrand, L., Pedersen, N. L. & Graham, D. Y. Helicobacter pylori infection: Genetic and environmental influences: a study of twins. Ann. Intern. Med. 120, 982–986 (1994).
9. Mayerle, J. et al. Identification of genetic loci associated with *Helicobacter pylori* serologic status. JAMA 309, 1912–1920 (2013).
10. Hansson, G. K. & Edfeldt, K. Toll to be paid at the gateway to the vessel wall. Arterioscler. Thromb. Vasc. Biol. 25, 1085–1087 (2005).
11. Lagunes-Servin, H. et al. Toll-like receptors and cytokines are upregulated during *Helicobacter pylori* infection in Children. Helicobacter 18, 423–432 (2013).
Y. L. and C. N. designed the study, D. X. H. and Y. Q. L. analysed the data. D. X. H., Y. Q. L., D. M. W. and Y. L. interpreted the data and wrote the manuscript. All authors read and approved the final version of the manuscript.

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

Acknowledgements
We thank for the funding support from Shenzhen Municipal Government of China (No. JCYJ20170412153155228).

Author contributions
Y. L. and C. N. designed the study, D. X. H. and Y. Q. L. analysed the data. D. X. H., Y. Q. L., D. M. W. and Y. L. interpreted the data and wrote the manuscript. All authors read and approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97790-7.

Correspondence and requests for materials should be addressed to C. N. or Y. L.

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

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

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

© The Author(s) 2021