

Reviewed Preprint

v1 • August 27, 2026

Not revised

For correspondence:

sdsb1@leicester.ac.uk

These authors contributed equally

& These authors contributed equally

Competing interests: No competing interests declared

Funding: See [page 21](#)

Reviewing editor: Detlef Weigel, Max Planck Institute for Biology Tübingen, Germany

© 2026, Archampong et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Genomic variation and ancestry of *Helicobacter pylori* in the admixed population of Cabo Verde reveal host adaptation, limited host–pathogen ancestry concordance, and signatures of trans-Atlantic slave trade migrations

Timothy Archampong^{1,2,#}, Yi Ling Tam^{1,#}, Charlotte Davison^{1,#}, Isra Boughanmi¹, Sam Tallman^{1,3}, Milani Rodrigues⁴, Neia Tavares⁴, Kleyde Simone Santos⁴, Francisco Barbosa Amado⁵, Isabel Ramos⁴, Antonieta Martins⁴, Richard Haigh^{1,6}, Marco R Oggioni^{1,7}, Rui Manuel Ferreira⁸, Céu Figueiredo^{8,9}, Fabian Freund¹, Isabel Inês Araújo^{4,10,11,&}, Sandra Beleza^{1,12,&} 

¹Division of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom • ²Department of Medicine, University of Ghana Medical School, Accra, Ghana • ³Genomics England, London, United Kingdom • ⁴Faculty of Science and Technology, University of Cabo Verde (Uni-CV), Praia, Cabo Verde • ⁵Digestive Endoscopy and Medical Clinic; R. Abílio Macedo, Praia, Cabo Verde • ⁶Leicester Drug Discovery and Diagnostics, University of Leicester, Leicester, United Kingdom • ⁷Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy • ⁸i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal • ⁹Faculty of Medicine, University of Porto, Porto, Portugal • ¹⁰Ne hHealth research cenTer of Cabo Verde – NEST-CV, Uni-CV, Praia, Cabo Verde • ¹¹Global Health and Tropical Medicine, Associate Laboratory in Translation and Innovation Towards Global Health, Institute of Hygiene and Tropical Medicine, NOVA University Lisbon, Lisbon, Portugal • ¹²Division of Microbiology and Infection, University of Leicester, Leicester, United Kingdom

eLife Assessment

This **valuable** study characterizes the population structure and ancestry of *Helicobacter pylori* in Cabo Verde. Using a carefully sampled population-based cohort and paired host–bacterial genomic data, the authors identify distinct local lineages but with limited concordance between human and bacterial ancestry. The evidence is **solid** for the principal findings concerning bacterial diversity, ancestry and recent lineage expansion, while the support for the claim that this reflects host adaptation, enhanced transmission, reduced virulence or a mix thereof is weaker. The work has particular relevance for researchers interested in microbial population genomics, host–pathogen coevolution and the genetic consequences of human migration and admixture.

<https://doi.org/10.7554/eLife.112292.1.sa2>

Abstract

The genomic structure of the stomach-colonising, gastric cancer–associated bacterium *Helicobacter pylori* mirrors human evolutionary history, making it difficult to disentangle shared ancestry from true host–pathogen interactions driving disease risk. Recently admixed populations offer opportunities for new coevolutionary trajectories between *H. pylori* and its host that may disrupt parallel evolution. Here, we analyse host *H. pylori* IgG, blood serum biomarkers, and

paired host-bacterial genomic data from the general population (independent of dyspepsia symptoms) of Cabo Verde. Seropositivity is high (82.8%), as in neighbouring African populations. Serum biomarkers indicate that this unique cohort mainly comprises hosts with limited gastric inflammation. Genomic analyses show that Cabo Verdean *H. pylori* derive European (hspSWeurope) and West African (hspAfrica1WAFrica) ancestry. The delineation of a reference dataset that includes new strains from Ghana and Portugal, uncovered a distinct West–Central African cluster (hspAfrica1WCAfrica). This dataset allowed inference of within-Africa and within-Europe ancestries in Cabo Verdean and American strains that reflect historical migrations linked to European colonisation and the trans-Atlantic Slave Trade. European strains in Cabo Verde show additional population structure, including two low-diversity clusters. Approximate Bayesian Computation (ABC) inference indicates that the most common cluster represents a clonal expansion, likely driven by host adaptation leading to increased transmission within Cabo Verde. Notably, these strains lack known virulence genes and show significant divergence from gastric cancer strains. Finally, host and bacterial ancestry are uncoupled, indicating disruption of ancestral parallel evolution. Cabo Verde, therefore, provides a unique natural model for studying host–pathogen interactions underlying susceptibility to *H. pylori*.

Introduction

Helicobacter pylori is an obligate human-associated bacterium whose evolutionary history is intrinsically linked to humans' history of migration since their origins in Africa 100,000 years ago (Falush et al., 2003 [↗](#); Linz et al., 2007 [↗](#); Moodley et al., 2012 [↗](#)). Such codiversification has led to global geographic patterns of genetic diversity in *H. pylori* that are congruent with those of their host (Suerbaum and Josenhans 2007 [↗](#)). As a result, population structure in *H. pylori* can be partitioned into major populations (“hp”) and subpopulations (“hsp”) with distinct geographic distributions, and patterns of genetic variation within these groups have been used to infer the demographic history of their host (Wirth et al., 2004 [↗](#); Moodley et al., 2021 [↗](#); Suzuki et al., 2022 [↗](#)). Recently, the *Helicobacter pylori* Genome Project (HpGP) has generated over 1000 *H. pylori* high-quality genomes from 50 countries to refine the bacterium's worldwide population structure and evolution, and which comprises a valuable reference dataset to evaluate population history and ancestry of bacterial populations across different geographical regions (Thorell et al., 2023 [↗](#)).

Despite being a frequent commensal coloniser of the gastric mucosa, *H. pylori* also exhibit great pathogenic potential and are a major risk factor for gastric cancer (de Martel et al., 2020 [↗](#)). Several genes in *H. pylori* have been associated with increased risk of gastric cancer such as the *cagA* (cytotoxin-associated gene) pathogenicity island (cagPAI), *vacA* (vacuolating cytotoxin), and *babA* (Lewis blood group antigen-binding adhesin) (Yamaoka 2010 [↗](#); Linz et al., 2024 [↗](#)). The presence of one or more of these genes does not however necessarily lead to disease (Linz et al., 2024 [↗](#)). In fact, despite commonly possessing virulence genes, the majority of *H. pylori* colonisations result in lifelong asymptomatic carriage; only 1–5% of the *H. pylori*-positive individuals develop gastric cancer, depending on ethnicity and environmental factors (Malfertheiner et al., 2023 [↗](#)). This indicates that disease risk in *H. pylori* infection is driven by complex combinations of bacterial virulence, host genetic susceptibility, and environmental diversity.

Parallel evolution and correlated population structure between *H. pylori* and humans makes it difficult to distinguish which cooccurrence of human-bacterium variants are due to parallel ancestral lineages and which are responsible for real biological interactions underlying differences in susceptibility across host populations. One way to overcome this problem is to analyse human-*H. pylori* interactions in recently admixed populations. Admixture brings together diverse genetic backgrounds from both the host and the microbe, thereby creating novel host-bacterium combinations that are absent in non-admixed populations and disrupting the ancestral genetic correlation between *H. pylori* and its host. For instance, although the genetic ancestry of *H. pylori* was associated with the genetic ancestry of their host in patients from two Colombian

communities, African strains were more virulent in hosts with high Native American ancestry, explaining the difference in gastric cancer incidence between those communities (Kodaman et al., 2014 [↗](#)).

Recently admixed populations with different admixture histories comprise novel opportunities for new evolutionary and coevolutionary trajectories between *H. pylori* and its host (Thorell et al., 2017 [↗](#); Muñoz-Ramirez et al., 2021 [↗](#)). For instance, “New World” strains from North, Central and South America showed highest divergence from “Old World” strains at virulence genes; this variation might contribute to higher virulence potential of the American strains. The disruption of previous parallel ancestry in these populations seems to lead to more pathogenic outcomes, akin to that seen in Kodaman et al. (2014) [↗](#). However, to fully appreciate the relationship between human and *H. pylori* ancestries and genetic variation that gives rise gastric disease, this relationship must also be characterised in asymptomatic, non-susceptible, individuals.

Here, we evaluate *H. pylori* genomic variation, population structure, and host-*H. pylori* codiversification trajectories that have arisen in Cabo Verde, an Atlantic archipelago off the coast of West Africa. Colonized by the Portuguese in the mid-15th century, the Cabo Verdean population derive mixed ancestry from Europe and West Africa (Beleza et al., 2012 [↗](#); Beleza et al., 2013 [↗](#); Korunes et al., 2022 [↗](#); Laurent et al., 2023 [↗](#)). We hypothesized that the unique population history and patterns of admixture will have broken down the correlation between host and *H. pylori* ancestry and genetic variation, and that Cabo Verde can, therefore, be considered as a model to evaluate host-*H. pylori* interactions underlying disease risk.

At the outset of this study, no epidemiological surveillance of *H. pylori* infection had been conducted in the islands. Therefore, we recruited volunteers from the general population, independently of the presence of dyspeptic symptoms, and collected paired host and *H. pylori* genomic data. This integrated dataset enabled us to: (i) conduct an unbiased epidemiological assessment of *H. pylori* colonisation in Cabo Verde; (ii) assess *H. pylori* genetic variation associated with human migrations during the Trans-Atlantic Slave trade; (iii) characterise the between-host and population-level *H. pylori* genetic variation, ancestry and evolution in Cabo Verde; and iv) evaluate the relationship between host and *H. pylori* ancestries across the islands and assess the extent to which Cabo Verde constitutes a natural model for disentangling host-microbe coevolution and disease-related interactions.

Results

H. pylori epidemiology and association with gastric inflammation indicators in Cabo Verde

In total, 876 Cabo Verdeans of the general population (regardless of dyspeptic symptoms) were evaluated for of *H. pylori* IgG in blood serum (Table 1 [↗](#)). There is an over-representation of women in the cohort (73% of the sample); the age ranged from 23 to 93 years old (mean $\pm$ SD= 43 $\pm$ 13.2). While most individuals (65%) were classified as being from Santiago, our two-generation island descent criterion provided a comparable dataset for the remaining islands (Table 1 [↗](#)). The cohort contained related individuals: 51 pairs/trios of first-degree relatives (either parents/offspring or sibling pairs/trios), three pairs/trios of second-degree relationship (one grandparent/grandchildren trio, one half-siblings trio, and one aunt/niece pair), and 24 couples. There was no difference in seropositivity between related and unrelated groups. Nonetheless, statistical analyses excluded related individuals.

There is no association between *H. pylori* serostatus and island group (or by island), sex or education level. Individuals with ages >60 years were significantly less seropositive than the other age groups (Odds Ratio, OR: 0.5; 95%CI: 0.2-0.9; pvalue= 0.035). This is associated with lower IgG concentration specially for individuals >70 years old (>70 years, n=29, regression coefficient, coeff (S.E.) = -16.6 (8.1) μ g/L, p-value= 0.04; Table 1- figure supplement 1 [↗](#)). This suggests that age-

Table 1. Socio-demographic characteristics of participants in the epidemiological survey, in total sample and stratified by *H. pylori* serostatus.

	Serum obtained N	Seronegative N (%)	Seropositive N (%)
Total	876	157 (18)	719 (82)
Island:			
Santiago	569	97 (17)	472 (83)
Fogo	106	16 (15)	90 (85)
Northwest ¹	49	11 (22)	38 (78)
Other ²	152	33 (22)	119 (78)
Sex:			
Male	235	38 (16)	197 (84)
Female	639	118 (18)	521 (82)
Age group (in years):			
20-30	146	23 (16)	123 (84)
30-40	244	39 (16)	205 (84)
40-50	199	35 (18)	164 (82)
50-60	175	31 (18)	144 (82)
>60	103	27 (26)	76 (74)
Education level:			
Lower	392	65 (17)	326 (83)
Intermediate	293	54 (18)	239 (82)
Higher	188	37 (20)	151 (80)

¹ Northwest includes individuals from Santo Antão, São Vicente, and São Nicolau Islands. This division is based on geographic proximity, related history of colonization and human population structure (Beleza et al., 2012; Korunes et al., 2022; Laurent et al., 2023).

² Other includes individuals from other islands or individuals whose parents were born in different islands.

related decline in antibody levels or sero-reversion contributes to a lower seropositivity in older individuals. The overall seropositivity in Cabo Verde (calculated for individuals <70 years old) is 82.8% (95%CI: 80.0-85.6%).

Evaluation of the serum levels of the pepsinogen I (PgI), pepsinogen II (PgII) and the PgI/PgII ratio allowed us to indirectly assess gastric inflammation (individuals with more severe gastritis have elevated PgII levels) and gastric mucosal atrophy (PgI <30 µg/L and/or PgI/PgII ratio <3 indicate increased risk of atrophic gastritis) of the individuals (Figure 1 [\[10\]](#)) (Di Mario et al., 2006 [\[11\]](#); Talaranta-Keerie et al., 2010 [\[12\]](#); Kitamura et al., 2015 [\[13\]](#)). PgI and PgII levels were higher and PgI/PgII ratio lower in seropositive individuals (PgI: Regression Coefficient, coeff (S.E.) = 18.0 (4.3) µg/L, p-value = 2.5×10^{-5} ; PgII: coeff (S.E.) = 3.8 (0.6) µg/L, p-value = 5.2×10^{-11} ; PgI/PgI: coeff (S.E.) = -3.7 (0.64) units, p-value = 1.1×10^{-8} ; controlling for sex and age). PgI/PgII ratios <3 were found in 19 individuals and were associated both with lower levels of PgI (Kruskal-Wallis test, p-value = 9.2×10^{-9} ; all except two individuals had PgI <30 µg/L) and higher levels of PgII (Kruskal-Wallis test, p-value = 0.001). Overall, PgI/PgII ratio <3 was associated with *H. pylori* seropositivity (chi-square = 4.8, 1 degree of freedom, p-value = 0.03).

Given the weak serological evidence for significant gastric mucosal atrophy (19 out of 876, equalling 2%), and that 60% of the seropositive individuals have PgII levels in the same range as seronegative individuals (PgII < 10 µg/L; Figure 1C [\[10\]](#)) (Kitamura et al., 2015 [\[13\]](#)), we can conclude that the sampled cohort is comprised mainly by *H. pylori* seropositive with limited gastric inflammation.

Subcontinental ancestry in Cabo Verdean and American *H. pylori* reflect host migrations during the Trans-Atlantic slave trade

We sequenced all *H. pylori* colonies isolated from 177 hosts and calculated the number of pairwise Single Nucleotide Variants (SNVs) across all within-host core genomes. In 35 hosts, within-host pairwise SNV differences overlapped with the range of observed between hosts, indicating multiple colonization in ~20% of the individuals. Given this, we assembled a dataset of 220 unrelated *H. pylori* isolates from 177 individuals from Cabo Verde, along with one *H. pylori* isolate from Senegal. We then ran haplotype-based ChromoPainter/fineSTRUCTURE (Lawson et al., 2012 [\[14\]](#)) and genotype-based ADMIXTURE (Alexander et al., 2009 [\[15\]](#)) analyses with a comprehensive worldwide isolate dataset (including those from but not restricted to the HpGP dataset; Figure 2 – source data 1), and the Cabo Verdean isolates to further investigate population structure and ancestry in the *H. pylori* populations resulting from the Trans-Atlantic slave trade (Figure 2 [\[16\]](#); Figure 2 – supplement figures 1 – 4).

We first inferred the continental ancestry of Cabo Verdean *H. pylori* using a subset of worldwide strains (Figure 2- source data 1; Figure 2- supplement figure 1). Cabo Verdean strains were found to spread between African and European populations in principal component analysis (PCA) of the coancestry matrix obtained with fineSTRUCTURE. There was no detectable genetic contribution from hpAmerind, hspEAsia, hpAsia2 and hpAfrica2 to the Cabo Verdean bacterial population and, therefore, these reference strains were excluded from further analyses.

Next, to define geographic-localised reference groups, we considered as European and African references only those isolates from Europe or Africa whose ancestry was determined by PCA, k-means clustering and fineSTRUCTURE to be from their respective continents (Figure 2- supplement figure 2). In a preliminary fineSTRUCTURE analysis with Cabo Verdean and American isolates, African isolates from Sudan, Cameroon and Nigeria were found to be outlying in PC2 of the coancestry PCA and were, therefore, removed from subsequent analyses (Figure 2- supplement figure 3). The final analyses were performed with a total of 1165 isolates and 37,480 core SNVs with complete genotype data (Figure 2 [\[16\]](#); Figure 2 - figure supplement 4-5 and Figure 2 – source data 1).

In the coancestry PCA resulting from this fineSTRUCTURE analysis, the European reference samples formed a cline of three clusters on PC2 (Figure 2A [\[16\]](#)), consistent with previous findings from the HpGP (Thorell et al., 2023 [\[17\]](#)). Accordingly, they were classified into hspNEurope,

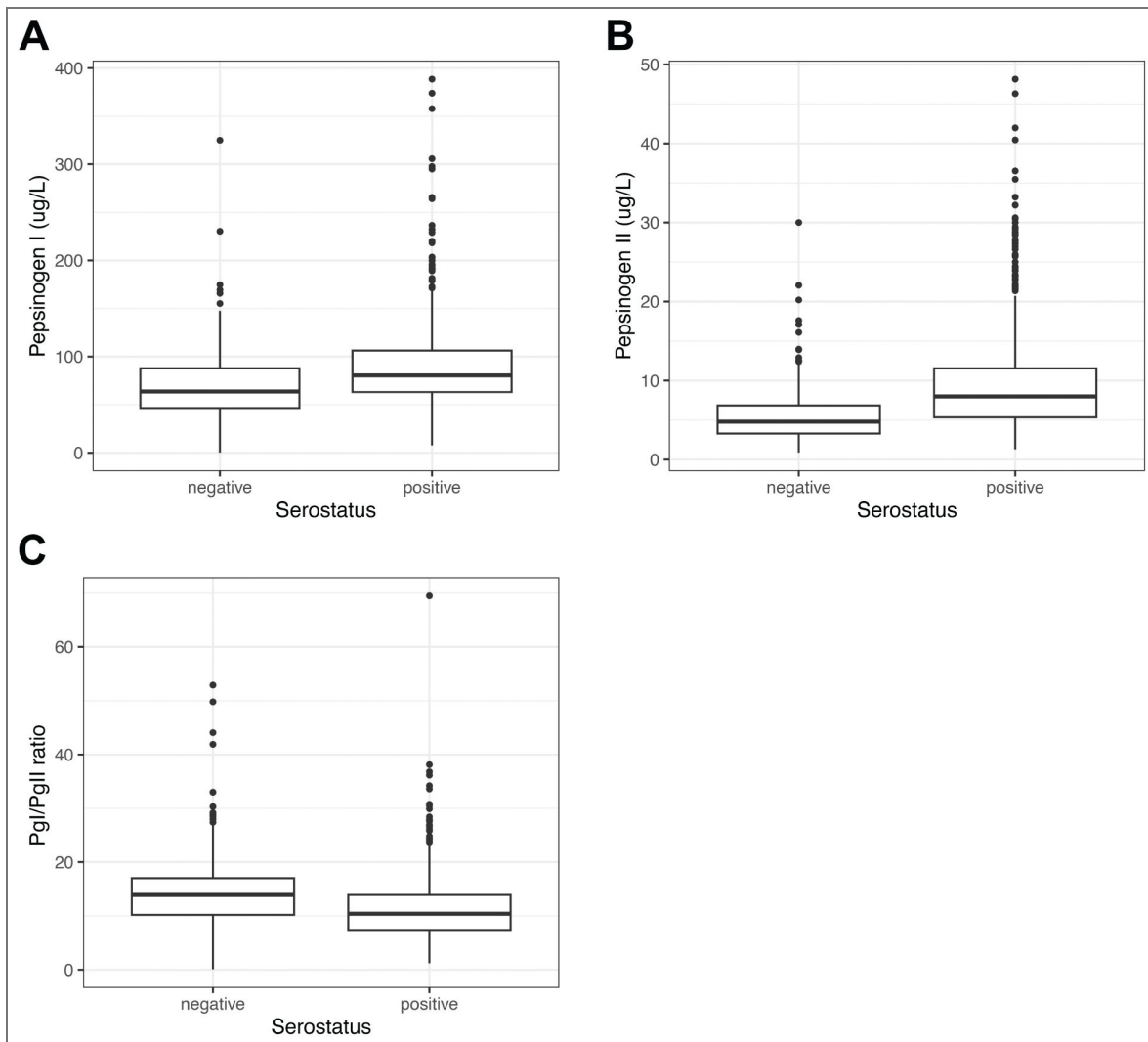
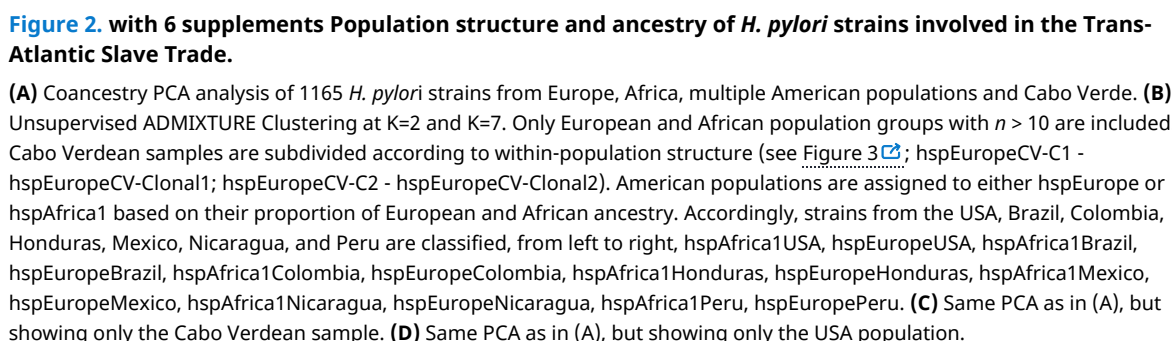


Figure 1. Distribution of pepsinogen I (A), pepsinogen II (B) and of the ratio pepsinogen I/pepsinogen II (PgI/PgII) (C) stratified by serostatus.



hspEurasia and hspSWEurope, following the HpGP classification scheme. This dataset includes 47 new genomes from Portugal from our own collection, all of which were classified as hspSWEurope.

In contrast, the addition of four new genomes from Ghana into the African reference dataset led to a new cline in PC2 (Figure 2A). These new Ghanaian isolates formed a new cluster together with four isolates from Ghana and Nigeria which have been previously classified as hspAfrica1NAmerica by the HpGP (Thorell et al., 2023). As this cluster is positioned between the previously described hspAfrica1WAfrica and hspAfrica1SAfrica and, in accordance with its geographic location in Africa, we refer to it here as hspAfrica1WestCentralAfrica (hspAfrica1WCAfrica).

These fineSTRUCTURE and ADMIXTURE (K=2-7) analyses allowed us to detect, for the first time, differences in subcontinental European and African ancestry among *H. pylori* genomes from Cabo Verde and 14 American countries (Figure 2; Figure 2- figure supplement Supplementary 4-5). On the African side, Cabo Verdean isolates traced their ancestry predominantly to hspAfrica1WAfrica, whereas most of the American isolates traced their ancestry to hspAfrica1WCAfrica (e.g. USA, Colombia, Dominican Republic, Nicaragua, Peru; with most of these having previously been classified as hspAfrica1NAmerica, Figure 2 – source data 1), or to hspAfrica1WCAfrica and hspAfrica1SAfrica (e.g. Brazil, Honduras). On the European side, most of the American and Cabo Verdean isolates appear to trace their ancestry to hspSWEurope, although some USA, Argentinian, and Brazilians isolates are also found to have ancestry to hspEurasia and hspNEurope.

Due to sample size for most American *H. pylori* populations being limited, it is not possible to draw definitive conclusions about their overall ancestral origins; however, it is noteworthy that the observed patterns align with the subcontinental African and European ancestry observed for their human hosts (Ongaro et al., 2019; Micheletti et al., 2020; Laurent et al., 2023).

There is additional population structure in Cabo Verdean *H. pylori*

The FineSTRUCTURE and ADMIXTURE (K=7) analyses reveal additional population structure in Cabo Verdean *H. pylori* that is not solely explained by ancestry (Figure 2A-C; Figure 2- figure supplement 4). PCA and phylogenetic analyses of Cabo Verdean *H. pylori* core genomes disclose four main clusters or clades (Figure 3A-B): two clusters closely related with hspAfrica1 and hspSWEurope (Figure 3 - figure supplement 1), and two additional clusters of isolates that form short-branched, star-like phylogenetic clades. Individual ancestries estimated using SOURCEFIND (Chacon-Duque et al., 2018) with hspSWEurope and hspAfrica1WAfrica as surrogates and each isolate as a target were congruent with ADMIXTURE (K=2; Figure 2B) ancestries (Spearman correlation = 0.88; Figure 3 -figure supplement 2). This analysis confirmed a major European ancestry component in these two unique Cabo Verdean groups. Estimation of population level ancestries allowing for self-copy with SOURCEFIND show that, unlike the references and other coexisting Cabo Verdean population groups, each of the two unique European-derived groups copy more haplotypes with each other than with any other population group (Figure 3C; Figure 3 – source data 1). In addition, the folded synonymous (neutral) site-frequency spectrum (ζ SFS, distribution minor-allele absolute frequency at synonymous variant sites) of these unique groups are skewed towards rare variation in comparison to the other coexisting Cabo Verdean population groups (Figure 3D). Collectively, these findings support that these two groups are comprised of closely related isolates resulting from recent clonal expansions. Following the nomenclature rules established by the HpGP but considering only continental-level ancestry of the isolates, the four Cabo Verdean *H. pylori* populations were termed hspAfrica1CaboVerde (n=103), hspEuropeCaboVerde (n=31), hspEuropeCaboVerde-Clonal1 (n=68) and hspEuropeCaboVerde-Clonal2 (n=19) (Figure 3).

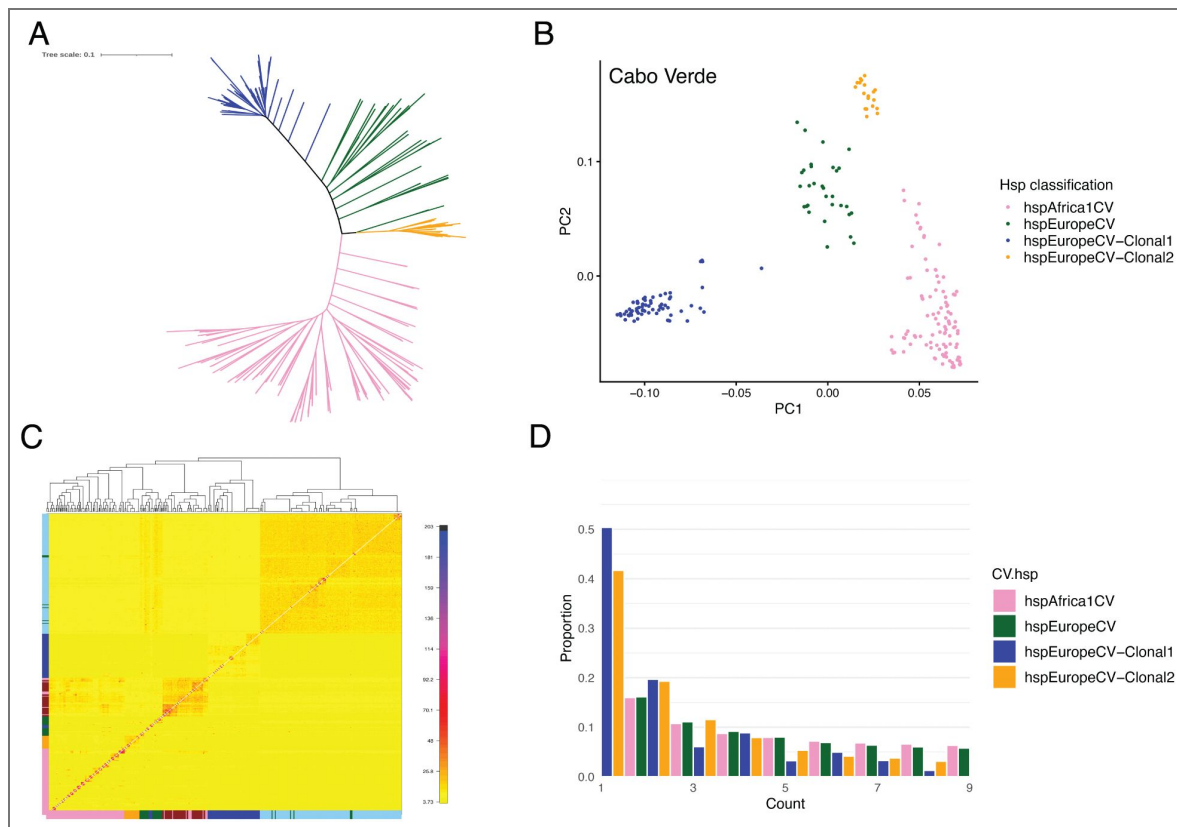


Figure 3. with 5 supplements Population Structure of *H. pylori* strain in Cabo Verde.

(A) Maximum likelihood unrooted phylogenetic tree of Cabo Verdean strains. Colour code is in (B) and (D). **(B)** PCA of Cabo Verdean strains. **(C)** ChromoPainter co-ancestry matrix with each cell indicating the proportion of DNA chunks that each genome copies from all other genomes (see Figure 3 – data source 1 for SOURCEFIND self-and between group copy fraction estimates). The colour bar on the left is repeated in the bottom of the matrix and denotes the Cabo Verdean population groups (with colour code in B and (D)), hspSWEurope (light blue) and hspAfrica1WAfrica (red). **(D)** Folded site frequency spectrum (fSFS) of synonymous variants in Cabo Verdean population groups. To get a comparable result, the fSFS was calculated based on the same sample size for all population groups.

Demographic inference supports increased population growth and transmission of hspEuropeCV-Clonal1 in Cabo Verde

We further investigated the demographic history of Cabo Verdean populations using the approximated Bayesian computation (ABC) approach based on the coalescent model from Davison et al. ((Davison et al., 2024 [DOI](#)); see methods), to infer a population growth model (exponential vs. no growth) and growth rate that best explain observed genetic variation in each Cabo Verdean population. Posterior predictive checks show that sample sizes do not affect model selection and only affect the growth rate estimation slightly (Figure 3 - figure supplement 3). Although growth rates have wide confidence estimates, our approach shows support for meaningful exponential population growth in hspEuropeCV-Clonal1 (growth rate estimate=13.7; 95% CI=[6.2,20.3]), followed by weaker growth in hspEuropeCV (growth rate estimate=1.1 ; 95% CI=[0.5,2]), and minimal growth in hspAfrica1CV (growth rate estimate =0.2 ; 95% CI=[0.1,0.5]). Due to very limited sample size, we did not attempt to model the genetic diversity of hspEuropeCV-Clonal2 or infer its population growth rate.

It has been shown that selection, and especially positive selection, produce genomic signatures and coalescent genealogies analogous to those of population expansions in bacteria (Lapierre et al., 2016 [DOI](#)). Given that the host environment is shared among all Cabo Verdean population groups (see below), the higher growth rate observed in hspEuropeCV-Clonal1 can best be explained by adaptive evolution following its introduction into Cabo Verde. While posterior predictive checks for Tajima's D support an exponential growth model (Figure 3 - figure supplement 3), alternative evolutionary explanations cannot be excluded. For instance, multiple merger genealogies due to selection effects could show similar genetic diversity signals (Goldberg 2026 [DOI](#)). This type of genealogy would be potentially driven by increased transmission of these strains in Cabo Verde.

Clonal expansion hspEuropeCV-Clonal1 has reduced virulence potential

To identify and characterize local adaptations within the hspEuropeCV-Clonal1 and hspEuropeCV-Clonal2 lineages, we calculated the Population Branch Statistic (PBS) for each core-genome SNV and accessory gene (Yi et al., 2010 [DOI](#)). PBS measures hspEuropeCV-clonal lineage-specific allele-frequency changes at a given locus relative to the divergence from hspEuropeCV and hspAfrica1CV, effectively identifying loci that have experienced rapid evolution since the divergence of these Cabo Verdean clonal groups.

Fold enrichment analysis of functional annotations for the most highly differentiated genes assigned based on eggNOG ortholog annotations (Huerta-Cepas et al., 2019 [DOI](#)) revealed a significant enrichment of core and accessory genes associated with cell wall and cell membrane biogenesis (COG category M) within hspEuropeCV-Clonal1 (Figure 4A [DOI](#); Figure 4 – source data 1). Due to limited sample size, hspEuropeCV-Clonal2 did not show significant enrichment of any core or accessory COG categories (Figure 4B [DOI](#); Figure 4 – source data 2).

Further clues to the genomic basis of lineage divergence and population expansion of hspEuropeCV-Clonal1 emerged from the analysis of accessory and virulence genes (Figure 4C [DOI](#), Figure 4 - source data 3). In contrast to hspEuropeCV-Clonal2, the hspEuropeCV-Clonal1 isolates form a distinct clade in a tree constructed from genetic distances calculated based on accessory gene presence/absence (Figure 4C [DOI](#)). Inspection of this tree reveals a cluster of genes that are largely absent in hspEuropeCV-Clonal1 but common in the remaining Cabo Verdean population groups (boxed region in Figure 4C [DOI](#)). These genes do not reach significant divergent PBS values due to their intermediate frequency in the parental hspEuropeCV (Figure 4–source data 3). Notably, this cluster corresponds to genes of the *cag* pathogenicity island (*cagPAI*), including *cagA*. An inspection of other virulent genes shows that, unlike the other population groups in Cabo Verde, *vacA* in hspEuropeCV-Clonal1 is of the less virulent type s2/i2/m2 (Figure 3d [DOI](#)). Finally, human hosts carrying hspEuropeCV-Clonal1 alone (n=56) tend to have higher pepsinogen ratios than hosts carrying the other lineages (n=112) (Mann-Whitney U = 3865.5, p-value = 0.0103).

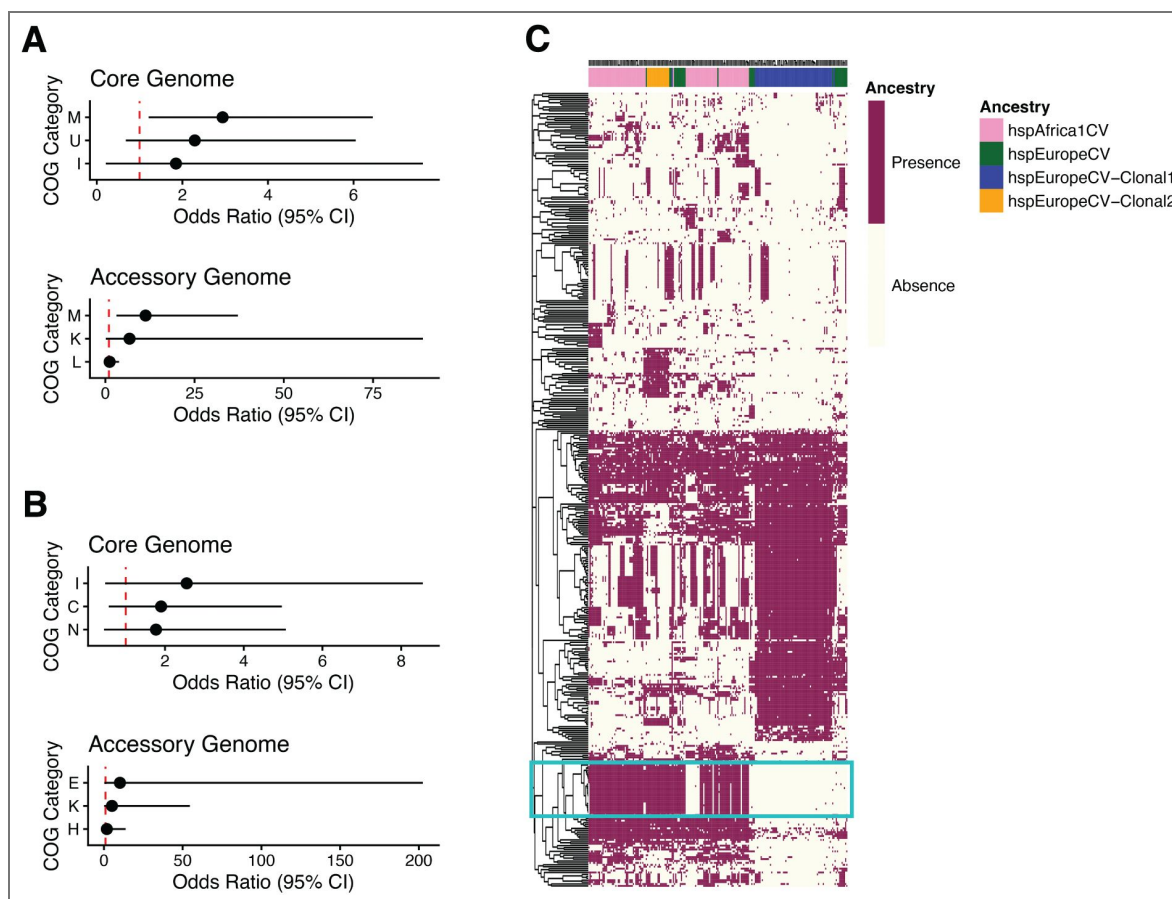


Figure 4. with 5 supplements Genomic composition of Cabo Verdean *H. pylori* population groups.

(A) Forest plots showing odds ratios and 95% confidence intervals for the top three genomic annotations in the core and accessory genomes, derived from a fold enrichment analysis of hspEuropeCV-Clonal1. (B) Forest plots showing odds ratios and 95% confidence intervals for the top three genomic annotations in the core and accessory genomes, derived from a fold enrichment analysis of hspEuropeCV-Clonal2. (C) Hierarchical clustering based on the accessory genomes' presence (burgundy)/absence (white) for the whole Cabo Verdean sample. Strains are coloured based on their population group on the top of the chart. Boxed region corresponds to *cagPAI*. Core genetic variants were called via mapping to the reference sequence 26695 and correspond to those present in $\geq 95\%$ of the total sample. Accessory genes were obtained through pangenome alignment and correspond to genes present in 5-95% of the total sample. COG category descriptions are in Figure 4 – source data 1-2.

Given that the Cabo Verdean strains were isolated from hosts with limited gastric inflammation, we carried out a genome-wide association analysis (GWAS) of gastric cancer (GCA) using the (Berthenet et al., 2018 [\[1\]](#)) dataset (Figure 4 – source data 4). To this end we employed the compaction of overlapping k-mers (termed unitigs) approach in DBGWAS (Jaillard et al., 2018 [\[2\]](#)), and included the Cabo Verdean European isolates as control groups.

The analysis replicated several hits reported by Berthenet et al. (2018) and showed that the majority of genes significantly enriched in GCA patients relative to controls correspond to well-established *H. pylori* virulence factors. These included genes within the *cagPAI* including *cagA*, genes encoding membrane adhesion proteins (*babA2* and *sabA*) and the vacuolating cytotoxin *vacA* (Figure 4 – source data 5). Around half of the gene hits present (66% unitigs; 52% of unique genes) showed frequency differences greater than 50% between GCA strains and hspEuropeCV-Clonal1. Within these hits, there was enrichment for genes involved in lipopolysaccharide and peptidoglycan biosynthesis (COG category M, twice the frequency of the next most represented COG categories P and U; Figure 4 – source data 5).

A notable observation was that 19% (n=37) of the unitigs present in 98% of hspEuropeCV-Clonal1 strains aligned to a non-coding region upstream of a multidrug efflux system ATP-binding protein (C694_RS06300 locus tag in reference sequence 26695 with NCBI reference NC018939.1; (Alfaray et al., 2023 [\[3\]](#))). The DBGWAS authors suggest that this signal may indicate the presence of a putative mobile genetic element in this region (Jaillard et al., 2018 [\[2\]](#)). However, searches in ICEfinder (Wang et al., 2024 [\[4\]](#)) and ISfinder (Siguier et al., 2006 [\[5\]](#)) databases, and prophage analysis with geNomad (Camargo et al., 2024 [\[6\]](#)) and PhageBoost (Siren et al., 2021 [\[7\]](#)) did not identify any integrative conjugative elements (ICEs), insertion sequences (ISs), or prophage elements in this genomic region. Further investigation will therefore be required to determine whether smaller, or highly diverged mobile genetic elements contribute to the observed signal in this genomic region.

Collectively, these analyses demonstrate that hspEuropeCV-Clonal1 exhibits a distinctive carriage-associated genetic background consistent with reduced virulence potential.

***H. pylori*-Host codiversification in Cabo Verde**

We estimated the individual African ancestry of the human hosts using ADMIXTURE (K=2) with the YRI (Yoruba from Ibadan) and IBR (Iberian from Spain) from the 1000 Genomes Project (Genomes Project et al., 2015) as reference panels. The median human West African ancestry in this cohort is 0.70 (range 0.47-0.83), which is within the range previously reported for the Santiago island (based on an independent genomic dataset for 685 Cabo Verdeans across the archipelago, median West African ancestry in Santiago: 0.74, range 0.45-0.86; higher than the median West African ancestry for the total archipelago: 0.58, range 0.24-0.88) (Beleza et al., 2013 [\[8\]](#); Korunes et al., 2022 [\[9\]](#)).

In this human genomic context, analysis of the relationship between human and infecting *H. pylori* genomic ancestry and structure in Cabo Verde was performed for each *H. pylori* population group. As individual African ancestry in *H. pylori*, we used the estimates obtained with SOURCEFIND because it provides more accurate admixture proportions by correcting for biases introduced by uneven sample sizes among the reference populations (Chacon-Duque et al., 2018 [\[10\]](#)).

Notably, the proportion of human West African ancestry was not correlated with the proportion of hspAfrica1WAfrica ancestry in any of the *H. pylori* population groups (Spearman's rho= 0.115; p-value= 0.096; Figure 5 [\[11\]](#)).

Discussion

Our analysis of *H. pylori* in Cabo Verde leverages a unique population sample shaped by founder effects and five centuries of European and African admixture. A major strength of our study lies in the use of samples from hosts with limited gastric inflammation, which offered two primary advantages: firstly, it allowed for an unbiased calculation of seroprevalence and, secondly, it

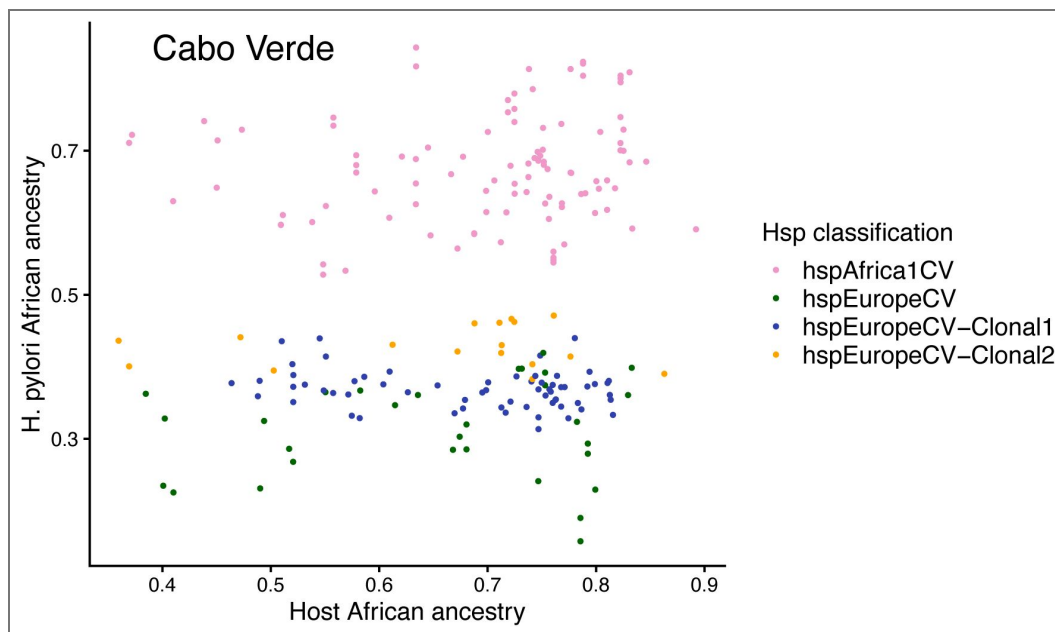


Figure 5. Host-*H. pylori* codiversification in Cabo Verde.

Host west African (YRI) ancestry (X-axis) is plotted against *H. pylori* African (hpAfrica1WestAfrica) ancestry (Y-axis).

provided a more representative sample of *H. pylori* genetic diversity and host–bacterium combinations within the archipelago to evaluate fully the population structure and evolution of *H. pylori* in the islands.

The estimated *H. pylori* prevalence of 83% (95%CI: 80.0–85.6%) in Cabo Verde (recorded in 2016) is comparable to rates observed in other African nations between 1989 and 2008 (Hooi et al., 2017 [↗](#); Zamani et al., 2018 [↗](#)). However, this figure is slightly higher than the mean prevalences reported for several African countries in a large meta-analysis covering the decade 2011–2022 (Li et al., 2023 [↗](#)). This discrepancy may be partially explained by observation of an overall decline in the prevalence of *H. pylori* in Africa (Zamani et al., 2018 [↗](#); Li et al., 2023 [↗](#)); this trend may not be evident in Cabo Verde due to the absence of longitudinal surveillance data. Additionally, it may also reflect our reliance on a serological method (which aligns with Hoi et al. (2017) and Zamani et al. (2018) but is in contrast with Li et al. (2013)), which does not distinguish active from past infections and tends to lead to higher prevalence estimates (Li et al., 2023 [↗](#)). The widespread colonization of people in Cabo Verde, regardless of sex, age and educational levels, highlights a pressing need for further study on risk variants for gastric disease within populations. As we have not evaluated diseased individuals here, future research should investigate whether risk profiles vary among admixed individuals based on their proportion of African ancestry.

Our haplotype- and genotype-based analyses refine our understanding of how the Trans-Atlantic slave trade shaped the subcontinental ancestry of *H. pylori* populations in Cabo Verde and the Americas. By integrating an expanded population reference dataset and restricting African and European source populations to genetically well-resolved lineages, we uncover new patterns of genome variation and admixture. In particular, the identification of a new distinct West–Central African cluster (hspAfrica1WCAfrica) highlights the genetic heterogeneity within African-derived *H. pylori* lineages and provides a parsimonious explanation for the ancestry of many American isolates formerly grouped under hspAfrica1NAmerica (Thorell et al., 2023 [↗](#)). In addition, the within-Africa and within-Europe ancestry patterns observed between Cabo Verdean and American strains are consistent with European colonisation and the historical origins of enslaved populations reaching these regions (Ongaro et al., 2019 [↗](#); Micheletti et al., 2020 [↗](#)). Together, these findings underscore the need to expand sampling of African *H. pylori* populations and to more comprehensively characterise genetic variation across Africa. Such efforts will allow refinement in the reconstruction of the evolutionary and historical trajectories of *H. pylori* in post-colonial populations. Importantly, they will also be critical for dissecting patterns of genetic variation underlying disease risk, both within Africa and in African-derived populations.

Genetic variation in the human population of Cabo Verde is predominantly explained by admixture (Korunes et al., 2022 [↗](#)). In contrast, ancestry-based population structure is not the main axis of genetic differentiation in the Cabo Verdean *H. pylori*. Instead, the major feature driving population structure in Cabo Verde is the presence of a lineage that most likely arose through a combination of genetic drift and admixture, which then subsequently expanded in the population through adaptive evolution. Notably, this lineage derives from a European ancestral background and has evolved towards a carriage-associated state (limited virulence compared with hspEuropeCV), showing significant divergence from gastric cancer strains. More work is needed to determine which of rapidly evolving candidate genes in the HspEuropeCV-Clonal1 lineage, relative to hspEuropeCV and hspAfrica1CV, contribute to its higher transmission and expansion in Cabo Verde.

Frequent cross-infection between unrelated individuals, admixture and clonal expansions have broken the correlation between host and *H. pylori* ancestries in the largely asymptomatic population of Cabo Verde. This contrasts with findings from symptomatic individuals in Colombia (Kodaman et al., 2014 [↗](#)), where a positive ancestry correlation between *H. pylori* and human host remains. However, to fully contrast with Kodaman et al. (2014 [↗](#)), we would need to evaluate if this relationship is similarly disrupted in diseased Cabo Verdean individuals. More broadly, the discrepancy in ancestry correlation between studies highlights the stochastic nature of admixture and its role in generating distinct coevolutionary trajectories that can lead to different phenotypic outcomes. The disrupted host-*H. pylori* ancestry correlation in Cabo Verde indicates that

admixture has been extensive enough to decouple the co-occurrence of specific human-*H. pylori* variants that are due to parallel ancestral lineages. As such, Cabo Verde represents a valuable natural system for investigating how host-*H. pylori* genetic interactions influence disease risk and progression.

Material and Methods

Ethics

The study was approved by the National Ethical Committee for Health Research of Cabo Verde (Deliberation n°19/2014) and by the Ethics Sub-Committee for Medicine and Biological Sciences of the University of Leicester (protocol 4996-sdsb1-genetics). All participants provided written consent after an explanation of the goal and procedures of the study. In the global population analysis, we additionally incorporated newly generated genome-wide sequences from four Ghanaian isolates obtained from patients presenting upper gastrointestinal symptoms, and 47 genome-wide sequences of Portuguese isolates from patients with gastritis. All isolates were obtained with participants' informed consent and under ethics approval by the Department of Medicine, Korle-Bu Teaching Hospital, Accra, Ghana, and by the University Hospital Centre (CHU), Porto, Portugal.

Sample collection, determination of *H. pylori* serostatus and indirect measurement of mucosal damage

Samples were collected in the capital city Praia, Santiago Island, Cabo Verde from February 2016 to April 2017. A participant was considered of Cabo Verdean origin only if the participant and their parents were native from Cabo Verde. All individuals of both sexes and >20 years of age were invited to participate regardless of showing *H. pylori* infection-related symptoms. Individuals that had self-reported use of antibiotics in the six months prior to sampling were excluded. All individuals provided saliva, blood and gastric mucus samples. Standard questionnaires were used to collect personal and family demographic data (age, sex, place of birth, education level) and *H. pylori* related symptoms (dyspeptic symptoms such as epigastric pain, nausea, and vomiting). An individual's island origin was defined based on their and their parents place of birth.

H. pylori status was determined by testing for IgG against *H. pylori* in the serum of centrifuged samples, using the human anti-*Helicobacter pylori* IgG ELISA Kit (Biohit Healthcare, Finland). The kit reports a sensitivity of 91% and a specificity of 89%. *H. pylori* IgG concentration in serum was also calculated from comparison with a standard curve drawn with control samples included in the ELISA kit.

Indirect measurement of gastric mucosal inflammation and atrophy were assessed by ELISA quantification of the blood biomarkers pepsinogen I and II (PgI and PgII; Gastroview kit, Biohit Healthcare, Finland; [Telaranta-Keerie et al., 2010]). The kit defines cut-offs of PgI <30 µg/L and/or PgI/PgII ratio<3 as indicative of atrophic gastritis [Telaranta-Keerie et al., 2010]).

Epidemiological assessment

We used *H. pylori* serostatus to assess the prevalence of *H. pylori* colonization in Cabo Verde and association with covariates previously described as risk factors, such as age, sex, education level (de Martel and Parsonnet 2006; Wang et al., 2025). Whenever possible (age, education level – see below, pepsinogens-see below), we defined covariates as categorical and continuous variables, and tested associations between variables and *H. pylori* serostatus using logistic and linear regression. The results were consistent between the two regression models. All analyses were performed considering unrelated individuals (hosts) only.

Educational level was evaluated using categorical and continuous variables. The categorical variable considered three levels: lower education (participant did not receive any formal education or received elementary schooling only); intermediate education (participant received

secondary schooling); higher education (participant received higher vocational schooling, or university and/or postgraduate education). The continuous variable consisted in the number of years in school education.

To assess gastric inflammation and atrophic gastritis, the levels of PgI, PgII, and their ratio (PgI/PgII) were compared between seropositive and seronegative individuals by regression analysis. We also evaluated how cut-offs for atrophic gastritis defined by the kit were associated with *H. pylori* serostatus using the Kruskal-Wallis test for continuous variables and the Pearson's chi-square test for categorical variables. Statistical analyses were performed in R.

H. pylori isolation and culture

Gastric mucus samples were collected for *H. pylori* isolation with the minimally invasive “Gastric String Test”, based on the (Velapatiño et al., 2006 [↗](#)) protocol. In brief, a 90cm cotton fibre in a gelatine capsule was swallowed by the fasting participant, but with one of the ends of the fibre being taped onto the subject's cheek. Following one hour, the string was retrieved and the distal 30cm that were in contact with the stomach was placed in transport medium (Brain heart infusion (BHI) broth (Oxoid, UK), 20% glycerol (Thermo Fisher Scientific, UK), 1% Skirrow antibiotic supplement (11.5 µg/ml vancomycin, 5.8 µg/ml Trimethoprim, 0.3 µg/ml Polymyxin B; Oxoid, UK), and 5µg/ml Amphotericin B (Thermo Fisher Scientific), and frozen in dry ice.

During culture, string and transport medium were vortexed vigorously for three minutes and then plated onto two selective BHI blood agar plates containing 5% brain heart infusion agar (BHIA), 7% defibrinated horse blood (Thermo Fisher Scientific), 1% BD BBL IsoVitaleX enrichment media (Thermo Fisher Scientific), 1% Skirrow antibiotic supplement, plus one of two antibiotic mixes: 5 µg/mL Amphotericin B only; or 5 µg/mL Amphotericin B and 30µg/mL Bacitracin. Samples were incubated for up to 10 days at 37°C, in microaerophilic conditions established by GENbag-GENbox systems (bioMérieux France). All small transparent, catalase and urease positive colonies were expanded and confirmed to be *H. pylori* colonies by PCR amplification of a minimum of two and a maximum of three housekeeping genes according to PubMLST amplification primers and conditions (<https://pubmlst.org/organisms/helicobacter-pylori/primers> [↗](#)). Bacterial DNA was extracted using the phenol/chloroform method (Chen and Kuo 1993 [↗](#)), with a modified initial lysis buffer composed by 40 mM Tris-Acetate (pH 9.0), 20 mM sodium acetate (pH 9.0), 60 mM EDTA (pH 9.0,) 1% SDS, to efficiently inactivate *H. pylori* nucleases.

H. pylori genome sequencing, read mapping and SNV calling

H. pylori genomic DNA was paired-end sequenced (Illumina HiSeq 4000 at Oxford Genomics Centre) to a mean coverage of 80x. Raw FASTQ reads were trimmed and filtered with Trimmomatic 2.0 using software specifications (Bolger et al., 2014 [↗](#)), and quality-checked using FastQC 0.11.5 (Andrews 2010 [↗](#)). Paired reads were mapped to the *H. pylori* 26695 reference genome (NCBI Ref Seq accession: NC018939.1; assembly accession: CP003904.1), and SNVs were called using Snippy v4.6.0 (Seemann 2015). SNV annotation was performed with snpEFF (Cingolani et al., 2012 [↗](#)), via Snippy by adding the flag ‘-reference’ and providing the reference's Genebank file. Bcftools v1.2.0 (Danecek et al., 2021 [↗](#)) was then used to merge isolates' VCF files and to subset on biallelic SNVs with the flags ‘-m2 -M2 -v snps’. To identify core genome SNVs, we ran PLINK v1.90b3 (Chang et al., 2015 [↗](#)) with ‘-freq’ to extract the genomic positions of SNVs that were present in either 100% (for population and ancestry analyses) or 95% (for per SNV and per gene F_{ST} and PBS analyses) of chromosomes. The resulting list of core positions was then used to subset the VCF file with VCFtools v1.14.0 (Danecek et al., 2011 [↗](#)).

H. pylori de novo assembly

Trimmed, paired reads were assembled to contig level using SPAdes 3.12 with both mismatch repair and coverage threshold (set to auto) parameters implemented (Bankevich et al., 2012 [↗](#)). Contigs were further assembled into scaffolds using the “assembly improvement pipeline” from Sanger Pathogens (Page, De Silva, et al., 2016). Assembly quality was assessed via QUAST 4.3.

The resulting draft genomes were annotated and aligned using Prokka v1.14.6 (Seemann 2014 [↗](#)) and PIRATE v1.0.5 (Bayliss et al., 2019 [↗](#)), considering 50, 70, 90 and 95% sequencing identity. Subsequently, core genomes were built as a concatenation of genes identified to be present in 100% of input sequences using PIRATE's output files and python script 'create_pangenome_alignment.pl' (found in PIRATE Github's page <https://github.com/SionBayliss/PIRATE/tree/master/scripts> [↗](#)), with the parameters --dosage 1.25 -t 100 -I PIRATE.gene_families.ordered.tsv -f ./feature_sequences/.

***H. pylori* reference dataset and genome alignment**

Complete genomes of 1705 global *H. pylori* isolates were retrieved from public repositories (NCBI, Datadryad; last accessed December 2024) to investigate worldwide population structure and genetic ancestry of *H. pylori* populations (Figure 2 – source data 1). This comprehensive dataset comprises isolates whose population of affiliation has been assessed previously and have been classified into major populations (“hp”) and subpopulations (“hsp”), including those from the *Helicobacter pylori* Genome Project (HpGP; (Thorell et al., 2023 [↗](#))), as well as isolates whose population of affiliation has not been previously assessed, four new Ghanaian genomes and 47 new Portuguese genomes from our own collection.

PIRATE genome alignments were quality-checked using QUAST v4.3 (Gurevich et al., 2013 [↗](#)), and only assemblies with a N50 below 20kb were excluded; this single quality control (QC) threshold enabled the inclusion of a broader representation of certain South America countries, as well as all African genomes available up to December 2024. Sequence annotation and alignment of these sequences together with the Cabo Verde dataset was done using Prokka v1.14.6 and PIRATE v1.0.5 as described above. SNP calling for each alignment was performed using SNP-sites (Page, Taylor, et al., 2016).

A preliminary analysis of Single Nucleotide Variant (SNV) differences between pairs of aligned genomes with SNP-dists (Seemann 2021) showed the presence of clonal or closely related isolate pairs (characterized by an excess of isolates falling within the first percentile of the pairwise SNV difference distribution (<1000 pairwise SNV differences) and mostly representing duplicated samples in the database, within-host comparisons, or related individual comparisons). To enable a more robust analysis of deeper evolutionary relationships among *H. pylori* strains, we have cleared the dataset of these recent, clonal, relationships. Figure 2 – source data 1 includes only those isolates that passed QUAST QC, were determined to be unrelated (not clonal) and that have less than 2% missing genotypes.

Population genetics and ancestry analyses in a global context

We first defined a reference dataset comprising isolates known to belong to hpAfrica1, hpAfrica2, hpEurope, hpAsia2, hspEAsia, hspAmerind and confirmed to have originated from corresponding geographic regions. We then carried out k-means clustering with the KMeans algorithm in Python's scikit-learn package (Pedregosa et al., 2012 [↗](#))(scripts can be found in (Cave 2025 [↗](#))) and rounds of PCA using PLINK v1.9 to further resolve the continental and subcontinental ancestry of unclassified isolates (Figure 2 – figure supplement 2). For more computationally efficient population genetic analyses of African, European, American and Cabo Verde isolates, we subsetting the European dataset to a final total sample of 459 isolates. The resulting alignment length was 860,079 bp.

We used fineSTRUCTURE v4 to run the complete ChromoPainter/fineSTRUCTURE analysis pipeline (Lawson et al., 2012 [↗](#)). FineSTRUCTURE v4 was run with default parameters: 50,000 iterations of both the burn-in and Markov chain Monte Carlo (MCMC) model. In preliminary analyses, we included genetic variants with ≤ 2% missing genotype data, and carried out genetic imputation using SHAPEIT v2.r837 (Delaneau et al., 2008 [↗](#)) with parameters --states 500 --burn 10 --main 50 --prune 10. The 'makeuniformrecfile.pl' (part of the fineSTRUCTURE v4.1.1 package) was used to produce a uniform recombination map across *H. pylori* genomes. To analyse population structure in detail, we ran PCA on the coancestry matrix from the “chunkcounts.out” output file. The matrix was first normalised by setting the diagonal to the average of each row, then zero meaning by

subtracting the row means. Subsequent fineSTRUCTURE analyses (excluding outlying isolates in Figure 2 -figure supplement 3) were performed using only genetic variants with complete genotype data. The final fineSTRUCTURE analysis contained 1165 isolates and 37,480 SNVs (Figure 2 – source data 1).

ADMIXTURE v1.3.0 (Alexander et al., 2009 [↗](#)) was ran for values of K=2-9 in the same final dataset. We pruned the data to remove sites in high linkage disequilibrium (LD) using PLINK v1.90b3 with parameters `--indep-pairwise 50 10 0.1`. A total of 16,910 SNVs with no missing genotypes was included in the analysis.

Population genetics and ancestry analyses in Cabo Verde

A phylogenetic tree was built from PIRATE alignments using FastTree v2.1 (Price et al., 2010 [↗](#)) with the generalised time-reversible model of nucleotide evolution and the discrete gamma model with 20 rate categories for rescaling. The tree was visualised in iTOL v5 (Letunic and Bork 2021 [↗](#)). PCA was performed in PLINK v1.9.

For ancestry analyses, we ran fineSTRUCTURE V4.1.1 with PIRATE alignments incorporating hspAfrica1WAfrica and hspSWEurope as ancestral population references. This analysis included a total of 212,074 SNV with no missing genotypes. The resulting coancestry matrix from the “chunkcounts.out” output file was visualised as a heat map, using the authors’ R functions available at <https://people.maths.bris.ac.uk/~madjl/finestructure/finestructureR.html> [↗](#).

To calculate population and individual level ancestry in Cabo Verdean *H. pylori* while accounting for differences in the reference sample sizes, we ran SOURCEFIND v2 (Chacon-Duque et al., 2018 [↗](#)) with the “*.chunklengths.out” output file from fineSTRUCTURE and using hspSWEurope and hspAfrica1WAfrica as surrogates. As recipients, we either included the two reference populations and the four Cabo Verdean population groups (for population-level ancestry estimates), or each Cabo Verdean isolate individually (for individual-level ancestry estimates). In all SOURCEFIND runs, we fixed the truncated Poisson prior to four surrogate groups per individual, allowing up to eight groups to contribute ancestry at each MCMC iteration. We performed 2,000,000 iterations with a 50,000-step burn-in, sampling mixture coefficients every 5,000 iterations. Final ancestry proportions were calculated as the mean of all posterior samples.

Hierarchical clustering based on the accessory genomes’ presence /absence of genes was performed with the pheatmap v1.0.13 package in R v4.5.2, using the complete linkage method. Only genes present between 95% and 5% of the isolates (n=414), as determined from the PIRATE pangene alignment, were used in this analysis.

ABC inference

We modelled the demographic history of the four Cabo Verdean population groups using the ABC framework from (Davison et al., 2024 [↗](#)), with the following modifications. For each population group, we generated a reference table of 100,000 simulated SNP datasets with the same sample size and number of synonymous (neutral) SNVs, S , as the observed genome-wide (hspAfrica1CV, $S = 138,091$; hspEuropeCV, $S = 153,087$; hspEuropeCV-Clonal1, $S = 143,501$; hspAfrica1CV, $S = 138,545$), and whose SFS is derived from an analytical approach (Freund et al., 2023 [↗](#)) that incorporates the Kingman coalescence model with exponential growth characterized by a specified growth rate g (including no growth ($g = 0$), that is constant population sized). Specified growth rates ranged [0-15] for hspAfrica1CV and hspEuropeCV and [0-30] in hspEuropeCV-Clonal1, increasing by a step size of 0.05. We then compared the genetic diversity obtained from the simulations with the observed genetic diversity via the Random Forest ABC classifier (ABC-RF) from the *abcrf* R-package (Raynal et al., 2019 [↗](#)) to get the posterior growth rate parameter distributions. To evaluate the accuracy of the model the normalized mean absolute error NMAE, which averages $|\text{estimated value} - \text{true value}| / \text{true value}$ across all simulations, for all other true growth rates combined.

To assess whether the model fits the observed genetic diversity, we performed graphical posterior predictive checks by calculating the Tajima’s D of 1,500 simulated genomes, obtained with growth rates drawn uniformly from the full-range ABC-RF posterior quantiles at intervals of 0.025 (Figure

3-Figure supplement 3).

Per SNP and per gene F_{ST} , PBS, and COG-function enrichment analysis

We calculated pairwise F_{ST} values for each SNV within the core genome and for the presence/absence profiles of accessory genes. We utilized Hudson's F_{ST} estimator, as it is less sensitive to differences in sample size and SNV ascertainment schemes (Bhatia et al., 2013). The core SNV F_{ST} values were computed using the scikit-allel Python package (Alistair and Harding 2017), while the F_{ST} values for accessory genes were calculated using a custom implementation of the Hudson formula in R. Core SNVs were called via mapping to the reference sequence 26695 and correspond to those present in $\geq 95\%$ of the total sample. Accessory genes were obtained through PIRATE's pangenome alignment and correspond to genes present in 5-95% of the total sample.

We then used these F_{ST} s to calculate Population Branch Statistic (PBS) values as in Yi et al. (2010). PBS summarizes a three-way comparison of F_{ST} s between a focal group (hspEuropeCV-Clonal1 and hspEuropeCV-Clonal2), a closely related population (hspEuropeCV) and an outgroup (hspAfrica1CV). Within the core genome, we focused our analysis on the mean PBS of non-synonymous SNVs, as these variants are more likely to hold functional significance. A core gene was identified as significantly differentiated in the hspEuropeCV-Clonal1 and hspEuropeCV-Clonal2 lineages if it met two criteria: first, that the proportion of its non-synonymous SNVs falling within the top 5% of the PBS distribution exceeded 15% and 10%, respectively (95th percentile of proportions); and second, that its overall mean non-synonymous PBS value ranked within the top 5% of all mean non-synonymous PBS values (Figure 4 – source data 3-4). An accessory gene was identified as significantly differentiated in the hspEuropeCV-Clonal1 and hspEuropeCV-Clonal2 lineages if its PBS values ranked within the top 5% of all PBS values.

Next, we used EggNog v5.0 (Huerta-Cepas et al., 2019), to identify orthologous groups for each gene and assign them to specific Cluster of Orthologous groups (COG) functional categories. This classification enabled a fold enrichment analysis using Fisher's exact test to identify overrepresented COG functional categories within the sets of most significantly differentiated core and accessory genes.

Genome-wide association analyses of gastric cancer strains

We collected 53 gastric cancer strains from Bethernet et al. (2018; Figure 4 – source data 4) and used these as cases in a GWAS. Cabo Verdean European strains isolated from hosts with pepsinogen ratios >3 were used as controls. Host mean age was 42.1 years, age range 23 – 86 years, with 75.1 % female.

Associations between genetic variants in both the core and accessory genomes and the binary case-control trait were tested using a De Bruijn graph-based GWAS (DBGWAS) approach implemented in DBGWAS v0.5.3 (Jaillard et al., 2018). DBGWAS is a k-mer-based method that represents sequence variation using compact De Bruijn graphs (cDBGs), in which nodes correspond to DNA fragments of variable length generated by the compaction of overlapping k-mers (termed unitigs). This graph-based representation substantially reduces k-mer redundancy and captures genomic diversity of k-mers within the broader genomic context of the study population. DBGWAS tests for both locus-specific (unitig) and lineage effects on the phenotype using a linear mixed model (LMM), thereby accounting for population structure. DBGWAS was run with default settings and the “-SFF 200” flag which generates the first 200 most significant unitigs.

All unitig hits were analysed using BLAST v2.13.0 (Altschul et al., 1990) to identify corresponding genes. EggNog v5.0 (Huerta-Cepas et al., 2019) was subsequently used to assign each gene to orthologous groups and to classify them into COG functional categories.

We also ran the hspEuropeCV-Clonal1 control isolates' improved *de novo* assemblies through the ICEfinder v2.0 (Wang et al., 2024) and ISfinder (Siguier et al., 2006) databases to identify ICES and ISs that could be found in the non-coding region upstream of a multidrug efflux system ATP-

binding protein (C694RS06300 locus tag in reference sequence 26695 with NCBI reference NC018939.1).

To identify prophages in the same genomic region we followed the pipeline from (Cave 2025 [↗](#)). Briefly, the machine learning algorithms of geNomad v1.7.5 (Camargo et al., 2024 [↗](#)) and PhageBoost v0.1.7 (Siren et al., 2021 [↗](#)) were employed, using Prokka-annotated fasta files as inputs. CheckV v1.5 (Nayfach et al., 2021 [↗](#)) was subsequently applied to assess the quality of the identified putative prophage elements. Only elements longer than 20,000 bp, containing at least one viral gene, and with a CheckV-estimated completeness greater than 50% were retained as prophages.

Human DNA collection, extraction, and sequencing

Human DNA was obtained from saliva and blood samples. Saliva was collected with Oragene DNA 600 saliva kits (DNA Genotek, Canada). Saliva DNA was extracted from 0.5ml of sample with the PrepIT-L2P | PT-L2P reagent according to the manufacturer's protocol (DNA Genotek) and minor changes: RNA was degraded via introduction of a 30 minute, 37°C incubation with 10mg/ml RNaseA (Thermo Fisher Scientific) after initial DNA release incubation; ethanol volume was increased to 1.5X the supernatant volume for DNA precipitation; ethanol wash strength was increased to 85%; and final resuspension was done in 10mM Tris HCl pH8.5. Blood samples were collected by venepuncture using a Vacutainer® Plastic K2EDTA tube with Lavender BD Hemogard™ Closure 6ml 13x100mm and Safety-Lok™ Blood Collection Set with Pre-attached Holder 21G 0.75" 12" tubing (Becton Dickinson, UK). DNA was extracted via the phenol-chloroform method.

Human DNA samples were whole genome sequenced to 1x mean coverage (Illumina HiSeq 4000) and imputed by Gencove (New York, USA), using their imputation algorithms for low-pass sequencing to call variants with average imputation $r^2 > 0.9$.

Human population genetics and ancestry analysis

Biallelic VCFs were filtered based on linkage disequilibrium, missingness per marker, Hardy-Weinberg Equilibrium and minor allele frequency using PLINK 1.9 (PLINK flags: --indep-pairwise 50 5 0.8, --geno 0.1, --hwe 1e-6, --maf 0.1). Remaining variants (1,207,081 sites) were used to calculate PCA using PLINK 1.9 and host ancestry using FastNGS admix (Jorsboe et al., 2017 [↗](#)) with Yoruban (YRB) and Iberian (IBR) reference samples from the 1000 Genomes Project (Genomes Project et al., 2015).

Data availability

All *H. pylori* genomic datasets used here, including the new genomes sequenced in this study, are available under accession numbers provided in Figure2- Source Data 1. The 26695 reference sequence is available on NCBI under Ref Seq accession NC018939.1 and assembly accession CP003904.1.

Acknowledgements

This work was funded by: the Medical Research Council (MRC-UK), with grant number MR/M01987X/1 awarded to SB; the Biotechnology and Biological Sciences Research Council (BBSRC-UK) and the Midlands Integrative Biosciences Training Partnership (MIBTP), grant numbers BB/M01116X/1 awarded to YLT and CD, and BB/T00746X/1 awarded to IB; and a University of Leicester Doctoral Program Training Fellowship awarded to ST. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

This research used the ALICE High Performance Computing facility at the University of Leicester. We thank Dr Roxana Zamudio-Zea for her technical assistance with laboratory experiments, Dr Ebony Cave for bioinformatics support, and Isandra Pina for her support at the early stages

of sample collection. We are especially grateful to Dr. Dulce Dupret for providing invaluable information regarding gastric pathologies in Cabo Verde.

We also sincerely thank Dr. Agostinho Neto Hospital, Cabo Verde's Health Department, the Ponta D'Água Health Centre, the Tira-Chapéu Health Centre, and the University of Cabo Verde for providing the facilities and professional support during data collection and follow-up of the participants. Finally, we are deeply grateful to the Cabo Verdean population for their participation in this study.

Additional files

[Supplementary figures and tables.](#)

[Figure 2-source data 1.](#)

[Figure 4-source data 1.](#)

[Figure 4-source data 2.](#)

[Figure 4-source data 4.](#)

[Figure 4-source data 5.](#)

Additional information

Funding

Funder	Grant reference number	Author
UKRI Medical Research Council (MRC)	MR/M01987X/1	Sandra Beleza
UKRI Biotechnology and Biological Sciences Research Council (AFRC)	BB/M01116X/1	Yi Ling Tam Charlotte Davison
UKRI Biotechnology and Biological Sciences Research Council (AFRC)	BB/T00746X/1	Isra Boughanmi

Author ORCID iDs

Sandra Beleza: <https://orcid.org/0000-0002-7004-3431>

References

- Alexander DH**, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**:1655-1664 <https://doi.org/10.1101/gr.094052.109> | [PubMed](#)
- Alfaray RI**, Saruuljavkhlan B, Fauzia KA, Torres RC, Thorell K, Dewi SR, Kryukov KA, Matsumoto T, Akada J, Vilaichone RK, *et al.* (2023) Global Antimicrobial Resistance Gene Study of *Helicobacter pylori*: Comparison of Detection Tools, ARG and Efflux Pump Gene Analysis, Worldwide Epidemiological Distribution, and Information Related to the Antimicrobial-Resistant Phenotype. *Antibiotics (Basel)* **12** <https://doi.org/10.3390/antibiotics12071118> | [PubMed](#)
- Alistair M**, Harding N (2017) cggh/scikit-allel. Zenodo. version: v1.3.3 <https://doi.org/10.5281/zenodo.822784>
- Altschul SF**, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**:403-410 [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) | [PubMed](#)
- Andrews S** (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data.
- Bankevich A**, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prijbelski AD, *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**:455-477 <https://doi.org/10.1089/cmb.2012.0021> | [PubMed](#)

7. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ (2019) PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* **8** <https://doi.org/10.1093/gigascience/giz119> | PubMed
8. Beleza S, Campos J, Lopes J, Araujo II, Hoppfer Almada A, Correia e Silva A, Parra EJ, Rocha J (2012) The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies. *Plos One* **7**:e51103 <https://doi.org/10.1371/journal.pone.0051103> | PubMed
9. Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araujo II, Anderson TM, Vilhjalmsen BJ, et al. (2013) Genetic architecture of skin and eye color in an African-European admixed population. *Plos Genetics* **9**:e1003372 <https://doi.org/10.1371/journal.pgen.1003372> | PubMed
10. Berthenet E, Yahara K, Thorell K, Pascoe B, Meric G, Mikhail JM, Engstrand L, Enroth H, Burette A, Megraud F, et al. (2018) A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol* **16**:84 <https://doi.org/10.1186/s12915-018-0550-3> | PubMed
11. Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. *Genome Res* **23**:1514-1521 <https://doi.org/10.1101/gr.154831.113> | PubMed
12. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114-2120 <https://doi.org/10.1093/bioinformatics/btu170> | PubMed
13. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC (2024) Identification of mobile genetic elements with geNomad. *Nat Biotechnol* **42**:1303-1312 <https://doi.org/10.1038/s41587-023-01953-y> | PubMed
14. Cave E (2025) *Understanding the evolutionary history and dynamics of the gut bacteria Helicobacter pylori in Cape Verde* University of Leicester.
15. Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, Barquera R, Quinto-Sanchez M, Gomez-Valdes J, Everardo Martinez P, Villamil-Ramirez H, et al. (2018) Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun* **9**:5388 <https://doi.org/10.1038/s41467-018-07748-z> | PubMed
16. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**:7 <https://doi.org/10.1186/s13742-015-0047-8> | PubMed
17. Chen WP, Kuo TT (1993) A simple and rapid method for the preparation of gram-negative bacterial genomic DNA. *Nucleic Acids Res* **21**:2260 <https://doi.org/10.1093/nar/21.9.2260> | PubMed
18. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**:80-92 <https://doi.org/10.4161/fly.19695> | PubMed
19. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. (2011) The variant call format and VCFtools. *Bioinformatics* **27**:2156-2158 <https://doi.org/10.1093/bioinformatics/btr330> | PubMed
20. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10** <https://doi.org/10.1093/gigascience/giab008> | PubMed
21. Davison C, Tallman S, de Ste-Croix M, Antonio M, Oggioni MR, Kwambana-Adams B, Freund F, Beleza S (2024) Long-term evolution of *Streptococcus mitis* and *Streptococcus pneumoniae* leads to higher genetic diversity within rather than between human populations. *Plos Genetics* **20**:e1011317 <https://doi.org/10.1371/journal.pgen.1011317> | PubMed
22. de Martel C, Georges D, Bray F, Ferlay J, Clifford GM (2020) Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **8**:e180-e190 [https://doi.org/10.1016/S2214-109X\(19\)30488-7](https://doi.org/10.1016/S2214-109X(19)30488-7) | PubMed

23. de Martel C, Parsonnet J (2006) *Helicobacter pylori* infection and gender: a meta-analysis of population-based prevalence surveys. *Dig Dis Sci* **51**:2292-2301 <https://doi.org/10.1007/s10620-006-9210-5> | PubMed
24. Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**:540 <https://doi.org/10.1186/1471-2105-9-540> | PubMed
25. Di Mario F, Cavallaro LG, Moussa AM, Caruana P, Merli R, Maini A, Bertolini S, Dal Bo N, Rugge M, Cavestro GM, et al. (2006) Usefulness of serum pepsinogens in *Helicobacter pylori* chronic gastritis: Relationship with inflammation, activity, and density of the bacterium. *Digestive Diseases and Sciences* **51**:1791-1795 <https://doi.org/10.1007/s10620-006-9206-1> | PubMed
26. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**:1582-1585 <https://doi.org/10.1126/science.1080857> | PubMed
27. Freund F, Kerdoncuff E, Matuszewski S, Lapierre M, Hildebrandt M, Jensen JD, Ferretti L, Lambert A, Sackton TB, Achaz G (2023) Interpreting the pervasive observation of U-shaped Site Frequency Spectra. *Plos Genetics* **19**:e1010677 <https://doi.org/10.1371/journal.pgen.1010677> | PubMed
28. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. (2015) A global reference for human genetic variation. *Nature* **526**:68-74 <https://doi.org/10.1038/nature15393> | PubMed
29. Goldberg A (2026) Rare variation in malaria parasites biases population-genetic inference. *bioRxiv* <https://doi.org/10.64898/2026.01.09.698000> | PubMed
30. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072-1075 <https://doi.org/10.1093/bioinformatics/btt086> | PubMed
31. Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, Tanyingoh D, Malfertheiner P, Graham DY, Wong WWS, Wu JCY, et al. (2017) Global Prevalence of *Helicobacter pylori* Infection: Systematic Review and Meta-Analysis. *Gastroenterology* **153**:420-429 <https://doi.org/10.1053/j.gastro.2017.04.022> | PubMed
32. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**:D309-D314 <https://doi.org/10.1093/nar/gky1085> | PubMed
33. Jaillard M, Lima L, Tournoud M, Mahe P, van Belkum A, Lacroix V, Jacob L (2018) A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *Plos Genetics* **14**:e1007758 <https://doi.org/10.1371/journal.pgen.1007758> | PubMed
34. Jorsboe E, Hanghoj K, Albrechtsen A (2017) fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33**:3148-3150 <https://doi.org/10.1093/bioinformatics/btx474> | PubMed
35. Kitamura Y, Yoshihara M, Ito M, Boda T, Matsuo T, Kotachi T, Tanaka S, Chayama K (2015) Diagnosis of *Helicobacter pylori*-induced gastritis by serum pepsinogen levels. *Journal of Gastroenterology and Hepatology* **30**:1473-1477 <https://doi.org/10.1111/jgh.12987> | PubMed
36. Kodaman N, Pazos A, Schneider BG, Piazzuelo MB, Mera R, Sobota RS, Sicinschi LA, Shaffer CL, Romero-Gallo J, de Sablet T, et al. (2014) Human and *Helicobacter pylori* coevolution shapes the risk of gastric disease. *Proc Natl Acad Sci U S A* **111**:1455-1460 <https://doi.org/10.1073/pnas.1318093111> | PubMed
37. Korunes KL, Soares-Souza GB, Bobrek K, Tang H, Araujo II, Goldberg A, Beleza S (2022) Sex-biased admixture and assortative mating shape genetic variation and influence demographic inference in admixed Cabo Verdeans. *G3 (Bethesda)* **12** <https://doi.org/10.1093/g3journal/jkac183> | PubMed
38. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EP (2016) The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. *Mol Biol Evol* **33**:1711-1725 <https://doi.org/10.1093/molbev/msw048> | PubMed

39. **Laurent R**, Szpiech ZA, da Costa SS, Thouzeau V, Fortes-Lima CA, Dessarps-Freichay F, Lemee L, Utge J, Rosenberg NA, Baptista M, *et al.* (2023) A genetic and linguistic analysis of the admixture histories of the islands of Cabo Verde. *eLife* **12** <https://doi.org/10.7554/eLife.79827> | [PubMed](#)
40. **Lawson DJ**, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *Plos Genetics* **8**:e1002453 <https://doi.org/10.1371/journal.pgen.1002453> | [PubMed](#)
41. **Letunic I**, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**:W293-W296 <https://doi.org/10.1093/nar/gkab301> | [PubMed](#)
42. **Li Y**, Choi H, Leung K, Jiang F, Graham DY, Leung WK (2023) Global prevalence of *Helicobacter pylori* infection between 1980 and 2022: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol* **8**:553-564 [https://doi.org/10.1016/S2468-1253\(23\)00070-5](https://doi.org/10.1016/S2468-1253(23)00070-5) | [PubMed](#)
43. **Linz B**, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**:915-918 <https://doi.org/10.1038/nature05562> | [PubMed](#)
44. **Linz B**, Sticht H, Tegtmeier N, Backert S (2024) Cancer-associated SNPs in bacteria: lessons from *Helicobacter pylori*. *Trends Microbiol* **32**:847-857 <https://doi.org/10.1016/j.tim.2024.02.001> | [PubMed](#)
45. **Malfertheiner P**, Camargo MC, El-Omar E, Liou JM, Peek R, Schulz C, Smith SI, Suerbaum S (2023) *Helicobacter pylori* infection. *Nat Rev Dis Primers* **9**:19 <https://doi.org/10.1038/s41572-023-00431-8> | [PubMed](#)
46. **Micheletti SJ**, Bryc K, Ancona Esselmann SG, Freyman WA, Moreno ME, Poznik GD, Shastri AJ, Beleza S, Mountain JL (2020) Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am J Hum Genet* **107**:265-277 <https://doi.org/10.1016/j.ajhg.2020.06.012> | [PubMed](#)
47. **Moodley Y**, Brunelli A, Ghirotto S, Klyubin A, Maady AS, Tyne W, Munoz-Ramirez ZY, Zhou Z, Manica A, Linz B, *et al.* (2021) *Helicobacter pylori*'s historical journey through Siberia and the Americas. *Proc Natl Acad Sci U S A* **118** <https://doi.org/10.1073/pnas.2015523118> | [PubMed](#)
48. **Moodley Y**, Linz B, Bond RP, Nieuwoudt M, Soodyall H, Schlebusch CM, Bernhoft S, Hale J, Suerbaum S, Mugisha L, *et al.* (2012) Age of the association between *Helicobacter pylori* and man. *PLoS Pathog* **8**:e1002693 <https://doi.org/10.1371/journal.ppat.1002693> | [PubMed](#)
49. **Muñoz-Ramirez ZY**, Pascoe B, Mendez-Tenorio A, Mourkas E, Sandoval-Motta S, Perez-Perez G, Morgan DR, Dominguez RL, Ortiz-Princz D, Cavazza ME, *et al.* (2021) A 500-year tale of co-evolution, adaptation, and virulence: *Helicobacter pylori* in the Americas. *ISME J* **15**:78-92 <https://doi.org/10.1038/s41396-020-00758-0> | [PubMed](#)
50. **Nayfach S**, Camargo AP, Schulz F, Eloie-Fadrosch E, Roux S, Kyrpides NC (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**:578-585 <https://doi.org/10.1038/s41587-020-00774-7> | [PubMed](#)
51. **Ongaro L**, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, Gnechi-Ruscione GA, Alarcon-Riquelme ME, Patin E, Wangkumhang P, *et al.* (2019) The Genomic Impact of European Colonization of the Americas. *Curr Biol* **29**:3974-3986 e3974 <https://doi.org/10.1016/j.cub.2019.09.076> | [PubMed](#)
52. **Page AJ**, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, Otto TD, Keane JA (2016) Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microbial Genomics* **2** <https://doi.org/10.1099/mgen.0.000083> | [PubMed](#)
53. **Page AJ**, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**:e000056 <https://doi.org/10.1099/mgen.0.000056> | [PubMed](#)
54. **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, *et al.* (2012) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**:5 <https://doi.org/10.48550/arxiv.1201.0490>
55. **Price MN**, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *Plos One* **5**:e9490 <https://doi.org/10.1371/journal.pone.0009490> | [PubMed](#)

56. Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A (2019) ABC random forests for Bayesian parameter inference. *Bioinformatics* **35**:1720-1728 <https://doi.org/10.1093/bioinformatics/bty867> | PubMed
57. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068-2069 <https://doi.org/10.1093/bioinformatics/btu153> | PubMed
58. (2015) snippy: fast bacterial variant calling from NGS reads. GitHub. <https://github.com/tseemann/snippy>
59. (2021) snp-dists. GitHub. <https://github.com/tseemann/snp-dists>
60. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**:D32-36 <https://doi.org/10.1093/nar/gkj014> | PubMed
61. Siren K, Millard A, Petersen B, Gilbert MTP, Clokie MRJ, Sicheritz-Ponten T (2021) Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genom Bioinform* **3**:lqaa109 <https://doi.org/10.1093/nargab/lqaa109> | PubMed
62. Suerbaum S, Josenhans C (2007) *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat Rev Microbiol* **5**:441-452 <https://doi.org/10.1038/nrmicro1658> | PubMed
63. Suzuki R, Saitou N, Matsuari O, Shiota S, Matsumoto T, Akada J, Kinjo N, Kinjo F, Teruya K, Shimoji M, et al. (2022) *Helicobacter pylori* genomes reveal Paleolithic human migration to the east end of Asia. *iScience* **25**:104477 <https://doi.org/10.1016/j.isci.2022.104477> | PubMed
64. Talaranta-Keerie A, Kara R, Paloheim L, Harkonen M, Sipponen P (2010) Prevalence of undiagnosed advanced atrophic corpus gastritis in Finland: an observational study among 4,256 volunteers without specific complaints. *Scandinavian Journal of Gastroenterology* **45**:1036-1041 <https://doi.org/10.3109/00365521.2010.487918> | PubMed
65. Thorell K, Munoz-Ramirez ZY, Wang D, Sandoval-Motta S, Boscolo Agostini R, Ghirotto S, Torres RC, Hp GPRN, Falush D, Camargo MC, et al. (2023) The *Helicobacter pylori* Genome Project: insights into *H. pylori* population structure from analysis of a worldwide collection of complete genomes. *Nat Commun* **14**:8184 <https://doi.org/10.1038/s41467-023-43562-y> | PubMed
66. Thorell K, Yahara K, Berthenet E, Lawson DJ, Mikhail J, Kato I, Mendez A, Rizzato C, Bravo MM, Suzuki R, et al. (2017) Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. *Plos Genetics* **13**:e1006546 <https://doi.org/10.1371/journal.pgen.1006546> | PubMed
67. Velapatiño B, Balqui J, Gilman RH, Bussalleu A, Quino W, Finger SA, Santivaney L, Herrera P, Piscoya A, Valdivia J, et al. (2006) Validation of string test for diagnosis of *Helicobacter pylori* infections. *J Clin Microbiol* **44**:976-980 <https://doi.org/10.1128/JCM.44.3.976-980.2006> | PubMed
68. Wang H, Wang Z, Yu J, Jin J (2025) Social determinants of health and *Helicobacter pylori* infection prevalence: a systematic review and meta-analysis. *Front Public Health* **13**:1703158 <https://doi.org/10.3389/fpubh.2025.1703158> | PubMed
69. Wang M, Liu G, Liu M, Tai C, Deng Z, Song J, Ou HY (2024) ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Res* **52**:D732-D737 <https://doi.org/10.1093/nar/gkad935> | PubMed
70. Wirth T, Wang X, Linz B, Novick RP, Lum JK, Blaser M, Morelli G, Falush D, Achtman M (2004) Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proc Natl Acad Sci U S A* **101**:4746-4751 <https://doi.org/10.1073/pnas.0306629101> | PubMed
71. Yamaoka Y (2010) Mechanisms of disease: *Helicobacter pylori* virulence factors. *Nat Rev Gastroenterol Hepatol* **7**:629-641 <https://doi.org/10.1038/nrgastro.2010.154> | PubMed
72. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**:75-78 <https://doi.org/10.1126/science.1190371> | PubMed

73. **Zamani M**, Ebrahimitabar F, Zamani V, Miller WH, Alizadeh-Navaei R, Shokri-Shirvani J, Derakhshan MH (2018) Systematic review with meta-analysis: the worldwide prevalence of *Helicobacter pylori* infection. *Aliment Pharmacol Ther* **47**:868-876 <https://doi.org/10.1111/apt.14561> | PubMed
- Beleza S** (2026) *Helicobacter pylori* whole genomes from Cabo Verde, Ghana and Portugal. NCBI BioProject. ID PRJNA1472681 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1472681>
- Thorell K et al** (2017) Data from: Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. Dryad Digital Repository. <https://doi.org/10.5061/dryad.8qp4n>
- Ailloud et al** (2019) *Helicobacter pylori* genome evolution within the human stomach. NCBI BioProject. ID PRJNA490474 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA490474>
- Sheppard Lab** (2018) *Helicobacter pylori* from clinical gastric infection. NCBI BioProject. ID PRJNA395900 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA395900>
- Didelot X et al** (2013) Genomic evolution and transmission of *Helicobacter pylori*. NCBI BioProject. ID PRJEB4103 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB4103>
- Enroth et al** (2000) *Helicobacter pylori* and risk of gastric cancer: a case-control study. NCBI BioProject. ID PRJNA1136026 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1136026>
- Gutiérrez-Escobar AJ et al** (2021) *Helicobacter pylori* Genome sequencing - Peru. NCBI BioProject. ID PRJNA625094 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA625094>
- Hu L et al** (2023) Antibiotic Resistance of *H. pylori* in Central South China: Phenotype and Genotype Revealed by Short and Long Reads Based Whole Genome Sequencing. NCBI BioProject. ID PRJNA816422 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA816422/>
- Mannion A et al** (2021) MIT-*Helicobacter pylori* Genome sequencing and assembly. NCBI BioProject. ID PRJNA715181 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA715181>
- Muñoz-Ramírez ZY et al** (2017) *Helicobacter pylori* Genome sequencing and assembly. NCBI BioProject. ID PRJNA338771 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA338771>
- Muñoz-Ramírez ZY et al** (2021) Ancestry and local adaptation of *H. pylori* in the Americas. NCBI BioProject. ID PRJNA601302 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA601302>
- Karolinska Institutet** (2022) *Helicobacter pylori* Genome sequencing and assembly. NCBI BioProject. ID PRJNA479414 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA479414>
- Thorell K et al** (2023) *Helicobacter pylori* Genome Project (HpGP). NCBI BioProject. ID PRJNA529500 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA529500>
- Wilkinson D et al** (2022) Genomic diversity of *Helicobacter pylori* populations from different regions of the human stomach. NCBI BioProject. ID PRJNA787419 <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA787419>

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors are trying to characterize the sources of *H. pylori* in an island system. They find that, like the humans, the bacteria are admixed, but there is no correlation within the island between human ancestry and bacterial ancestry.

Strengths:

The study has taken particular care to characterize the humans from which isolates were obtained. Thus, it is a particularly convincing demonstration of a bacterial "melting pot".

Weaknesses:

The GWAS is highly confounded by population structure. In particular, there is a large group of strains that lack the *cag* pathogenicity island, and also differ in frequencies of other genes. So it's not clear that these differences are other than in *cag* status.

Figure 1 seems very inconclusive.

<https://doi.org/10.7554/eLife.112292.1.sa1>

Reviewer #2 (Public review):

Summary:

This study investigates the population structure and ancestry of *Helicobacter pylori* in Cabo Verde, where the human population has mixed West African and European ancestry. The authors combine a population survey, serum markers, bacterial genome analysis, and paired human-bacterial ancestry data. They report high *H. pylori* seropositivity, several distinct bacterial groups, and limited correlation between human ancestry and bacterial ancestry. They also identify one European-derived bacterial group that appears to have undergone a recent expansion and carries fewer well-known virulence-related genes.

The study is interesting, and the dataset is valuable, especially because population-based *H. pylori* genomic data from Cabo Verde and West Africa are limited. The results provide useful information on bacterial diversity, historical migration, and host-bacterial ancestry. However, some of the main conclusions are stronger than the evidence currently supports, particularly the claims of host adaptation, increased transmission, and reduced virulence.

Strengths:

- (1) A major strength is the study population. Participants were recruited from the general population and not only from patients with gastrointestinal disease. This gives a broader view of *H. pylori* diversity in Cabo Verde than studies based only on hospital patients.
- (2) The number of participants tested for *H. pylori* antibodies is substantial, and the authors also obtained a relatively large number of bacterial genomes. The combination of human and bacterial genomic information is another important strength. This allows the authors to directly examine whether human ancestry is related to the ancestry of the colonising bacteria.
- (3) The population genetic analyses are extensive. The authors use several different approaches, and these generally support the existence of African-derived and European-derived bacterial groups in Cabo Verde. The identification of two low-diversity European-derived groups is also interesting and suggests a relatively recent expansion.
- (4) The addition of new strains from Ghana and Portugal improves the reference dataset. The results may help future studies of *H. pylori* population structure in Africa, Europe, Cabo Verde, and populations affected by historical Atlantic migration.
- (5) The finding that human ancestry and bacterial ancestry are only weakly related in this population is potentially important. It suggests that the long-term relationship between human and bacterial ancestry may be less stable in recently admixed populations.

Weaknesses:

The main weakness is that several biological conclusions are based on indirect evidence. The genomic results support recent expansion of one bacterial group, but they do not directly show that this expansion was caused by adaptation to local hosts or by increased transmission. Founder effects, population history, geographic clustering, household

transmission, or random expansion may also explain the pattern. The wording should therefore be more cautious.

The conclusion of reduced virulence is also not fully supported. The expanded lineage often lacks the *cag* pathogenicity island and carries less virulent forms of *vacA*, which suggests lower virulence potential. However, this does not prove that the strains cause less gastric damage or lower disease risk. There are no endoscopic or histological data, and serum pepsinogen values are only indirect markers.

The description of the study population as having limited gastric inflammation is too strong. Serum pepsinogen measurements are useful for estimating gastric atrophy, but they do not directly measure the degree of histological gastritis. In addition, participants were recruited independently of symptoms, but this does not mean that they were all asymptomatic.

The epidemiological estimate is based on antibody testing. This measures seropositivity and cannot clearly distinguish current from previous infection. Therefore, terms such as active infection or colonisation should be used carefully.

The proposed new West-Central African bacterial group is based on a small number of reference strains from Ghana and Nigeria. The result is interesting, but broader sampling from African countries is needed before this group can be considered firmly established.

The interpretation related to the trans-Atlantic slave trade is plausible, but the data mainly show patterns consistent with known historical migration. They do not directly demonstrate when or how the bacterial lineages moved.

The gastric cancer comparison may also be affected by bacterial population structure. Differences between the Cabo Verdean lineage and gastric cancer strains may reflect ancestry or lineage differences rather than disease association alone.

Overall, the study achieves its main aim of describing *H. pylori* diversity and ancestry in Cabo Verde. The evidence is strong for the population structure and ancestry findings, but less strong for the proposed mechanisms of adaptation, transmission, and reduced disease-causing potential. The work will be useful to the field, but the main conclusions should be stated more carefully.

<https://doi.org/10.7554/eLife.112292.1.sa0>