

Helicobacter pylori Genome Aggregation Database reveals complex evolutionary forces shaping its genomic landscape and clinical impact in East Asia

Guanglin He

Guanglinhesu@163.com

Sichuan University <https://orcid.org/0000-0002-6614-5267>

Lintao Luo

Sichuan University <https://orcid.org/0009-0001-5790-4866>

Xiaoqiong Tang

Sichuan University

Hao Chen

Sichuan University

Yunhui Liu

<https://orcid.org/0009-0005-7252-5169>

Limei Zhang

Chongqing Medical University

Jie Zhong

Sichuan University <https://orcid.org/0000-0003-0354-9448>

Bowen Li

Sichuan University

Yutong Jiang

Chongqing Medical University

Zihao Zhu

Chongqing Medical University

Yuzhu Wang

Chongqing Medical University

Yuguo Huang

Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, China

Fengxiao Bu

West China Hospital, Sichuan University <https://orcid.org/0000-0002-4882-4856>

Huijun Yuan

The First Affiliated Hospital of Army Medical University

Haohao Dong

Sichuan University <https://orcid.org/0000-0002-5022-3494>

Binwu Ying

Sichuan University

Hong Li

Sichuan University

Renkuan Tang

Chongqing Medical University <https://orcid.org/0000-0003-0008-1288>

Mengge Wang

Sichuan University <https://orcid.org/0000-0002-3673-1855>

Chao Liu

Anti-Drug Technology Center of Guangdong Province <https://orcid.org/0000-0001-5633-3929>

Article

Keywords:

Posted Date: August 28th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7366252/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Helicobacter pylori genome aggregation database reveals complex evolutionary forces shaping its genomic landscape and clinical impact in East Asia

Lintao Luo^{1,2,3,4,#}, Xiaoqiong Tang^{5,6,#}, Hao Chen^{7,8,#}, Yunhui Liu^{1,2}, Limei Zhang^{1,2}, Jie Zhong^{3,4}, Bowen Li^{3,4}, Yutong Jiang^{1,2}, Zihao Zhu^{1,2}, Yuzhu Wang^{3,4,9}, Yuguo Huang^{3,4}, Fengxiao Bu⁴, Huijun Yuan^{3,4}, Chao Liu¹⁰, Haohao Dong^{7*}, Binwu Ying^{7,8,*}, Hong Li^{5,6,*}, Renkuan Tang^{1,2,*}, Mengge Wang^{1,2,*}, Guanglin He^{3,4,#,*}

¹Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China

²Chongqing Key Laboratory of Forensic Science & Chongqing Engineering Research Center of Criminal Investigation Technology, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China

³Center for Archaeological Science, Sichuan University, Sichuan 610000, China

⁴Department of Oto-Rhino-Laryngology, Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Sichuan 610000, China

⁵Center of Infectious Diseases, West China Hospital of Sichuan University, Sichuan 610000, China

⁶Laboratory of Infectious and Liver Diseases, Institute of Infectious Diseases, West China Hospital of Sichuan University, Sichuan 610000, China

⁷Department of Laboratory Medicine, State Key Laboratory of Biotherapy, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University and Collaborative Innovation Center of Biotherapy, Chengdu, Sichuan 610041, China

⁸Sichuan Clinical Research Center for Laboratory Medicine, Chengdu, Sichuan 610041, China

⁹Institute of Basic Medicine and Forensic Medicine, North Sichuan Medical College, Nanchong, Sichuan 637007, China

¹⁰Anti-Drug Technology Center of Guangdong Province, Guangdong 510640, China

#These authors contributed equally to this work

*Correspondence: Menggewang2021@163.com (M.W.); haohadong@scu.edu.cn (H.D.); lihong@wchscu.cn (H.L.); Renktang2012@163.com (R.T.); yingbinwu@scu.edu.cn (B.Y.); guanglinhesu@163.com (G.H.)

Abstract

We assembled HPgnomAD, a comprehensive global genome aggregation database of *Helicobacter pylori* (*H. pylori*) featuring 7,544 high-quality genomes from 278 populations, along with the first species-wide haplotype reference panel for genotype imputation. The panel provides high accuracy across diverse *H. pylori* populations, including those from low-coverage data, and is openly accessible with integrated analysis tools (<https://www.hpgnomad.top/>). Variant discovery revealed 1.82 million SNPs and 0.65 million InDels, with African strains exhibiting the most tremendous diversity and East Asian strains showing unexpectedly high novel variation. Fine-scale analysis of the hpEAsia lineage revealed six new sublineages shaped by altitude-, latitude-related divergence and region-specific admixture. Phylogenomic dating revealed two divergence waves in East Asia, paralleling Upper Paleolithic settlement and Neolithic human expansions, with highland-lowland separation at ~18.8 kya, followed by the formation of complex geography-related population substructures. Genome-wide scans revealed adaptive loci related to metal acquisition, nitrogen metabolism, surface adhesion, and membrane transport, including altitude-associated highly differentiated variants linked to antibiotic resistance and gastric disease. By integrating genomic, evolutionary, environmental, and clinical data, HPgnomAD offers a framework for understanding *H. pylori* evolution, host-pathogen adaptation, and precision medicine worldwide.

Keywords: HPgnomAD, *H. pylori* genomic diversity, Population structure, Adaptive evolution, Host-pathogen co-evolution

Introduction

Helicobacter pylori (*H. pylori*) predominantly colonizes the human stomach and has coexisted with humans for approximately 70,000 years; *H. pylori* is now known as a primary etiological agent of chronic inflammation, gastritis, peptic ulcers, and gastric malignancies¹⁻⁸. Its genome consists of a circular, double-stranded DNA molecule approximately 1.6–1.7 million base pairs in length⁹. Extensive recombination and high mutation rates have driven exceptional genetic diversity, nearly 50 times greater than that observed in humans^{10,11}. As a result, *H. pylori* strains display considerable genotypic and phenotypic variation across global populations¹²⁻¹⁶. Because *H. pylori* evolution closely mirrors human evolutionary history, its genome serves as a powerful proxy for reconstructing ancient human migrations^{14,17,18}. Large-scale human genomic data and ancient DNA studies further suggest a more complex evolutionary history for global *H. pylori* populations than previously assumed¹⁹⁻²³. Recent genomic studies have significantly advanced our understanding of the global distribution and phylogeographic structure of *H. pylori*, particularly through efforts such as the *H. pylori* Genome Project (HpGP), which systematically classified strains into distinct populations and subpopulations¹⁵. Building on this paradigm and a larger-scale *H. pylori* genomic dataset, Tournette *et al.* introduced the concept of "ecospecies" to reconcile discrepancies between genomic divergence and species definitions, incorporating large-scale data, including Indigenous American strains¹⁶. While such studies have generated invaluable resources for exploring *H. pylori* evolutionary dynamics, high-quality, population-wide databases capturing SNPs, InDels, and structural variants remain scarce.

Previous genetic studies have characterized the global population structure of *H. pylori*, revealing distinct clustering patterns and subpopulations^{15,24}. In East Asia, lineages such as hpEAsia, hpNorthAsia, and hpAsia2 appear to contribute significantly to regional population dynamics^{15,24}. Clinical epidemiological data have further highlighted the high prevalence of gastric cancer in East Asia, and with a differentiated occurrence rate among both highland and lowland East Asian populations²⁵⁻²⁸. Recently, the expansion of genomic resources from East Asian *H. pylori* strains has enabled a deeper investigation into their diversity and potential roles in the

pathogenesis of gastrointestinal disease^{29,30}. You *et al.* identified seven subregional clusters by analyzing 381 genomes and reported pronounced allelic variation at several loci, potentially linked to ecological adaptation²⁹. In parallel, Suzuki *et al.* analyzed 88 genomes from Okinawa, including 21 newly sequenced strains, to support a model of Paleolithic human migration from Africa to East Asia³⁰. Despite these advances, both studies present notable limitations in terms of the availability of high-quality databases, the scale of population stratification, and deep demographic reconstruction from a worldwide perspective. Most works are restricted to *H. pylori* isolate datasets, limiting the ability to draw continent-wide demographic and temporal inferences. Suzuki *et al.* relied on multi-locus sequence typing (MLST), which excludes genome-wide variation, potentially missing subtle yet biologically essential signals. While large-scale resources from the studies of HpGP and Elise *et al.* include East Asian *H. pylori* genomes^{15,16}, they provide limited insight into high-resolution genetic variation, fine-scale structure, biological adaptation, and historical relevance. Moreover, population-specific reference datasets for clinical and basic research remain insufficiently developed. The prevalence of *H. pylori* infection varies markedly across East Asia and is shaped by complex evolutionary pressures. These include extreme conditions, such as cold, hypoxia, and resource scarcity on the Qinghai-Xizang Plateau; distinct dietary practices across pastoral-agricultural regions of Northeast Asia; seafood-rich diets in southern East Asia; and heterogeneous pathogen exposures³¹⁻³⁵. Large-scale human genomic studies have identified adaptive signatures driven by these pressures^{32,36,37}. Additionally, human genetic modeling based on autosomal, Y-chromosomal, and mitochondrial data has revealed detailed population substructures linked to geography or language in East Asia and other regions^{19,21,33,38}. However, how complex human migration and admixture events affect the diversity and structure of geographically distinct *H. pylori* populations and the extent to which such evolutionary factors shape *H. pylori*'s genomic diversity and contribute to region-specific clinical outcomes remain unclear.

To address the limitations posed by the absence of a comprehensive genomic database for *H. pylori* and the scarcity of prior genomic diversity initiatives, we assembled a dataset of 8,010 high-quality *H. pylori* whole-genome sequences from both in-house sources and published studies^{12,15,29,39-62}. Using this dataset, we constructed the *H. pylori* Genome Aggregation Database (HPgnomAD) to investigate genomic diversity and variant spectra across diverse *H. pylori* lineages^{12,15,29,39-62} (**Fig. 1a; Extended Data Fig. 1, and Extended Data Table 1**). We employed multiple state-of-the-art population admixture modeling approaches, including model-based ADMIXTURE, fineSTRUCTURE, ChromoPainter, and BEAST-based phylogenetic reconstruction, to resolve the fine-scale population structure and infer the divergence times and expansion patterns of genetically distinct *H. pylori* lineages. In parallel, we applied various statistical methods (genome-wide association studies (GWAS), F_{ST} , and dN/dS analyses) to detect signals of natural selection among highland and lowland sublineages, as well as between northern and southern East Asian populations. These analyses revealed adaptive signatures with clinically relevant implications for phenotypic variation. To facilitate broader access, we developed an online platform (<https://www.hpgnomad.top/>) that catalogs the observed genomic variants, polymorphisms, and associated clinical features of *H. pylori*, providing a resource for researchers and clinicians. By integrating evolutionary and medical perspectives, we present a systematic characterization of the genomic architecture and adaptive trajectories of *H. pylori* in East Asia. Collectively, these findings refine our understanding of host-pathogen co-evolution and offer insights into population-specific prevention and control strategies.

Results

HPgnomAD construction, haplotype reference panel reconstruction, and variant discovery

We compiled high-quality *H. pylori* whole-genome sequences from in-house resources, published studies^{15,16,29,30,51,58}, and public repositories, including Enterobase and the National Library of Medicine. This effort produced the HPgnomAD resource, comprising 8,010 genomes from 278 global sampling sites. After stringent quality control, we curated a robust dataset of 7,544 genomes (7,542 *H. pylori* and 2 *H. acinonychis*) for foundational population genetic and evolutionary research (**Fig. 1a; Extended Data Fig. 1a**). From HPgnomAD, we constructed an *H. pylori* haplotype reference panel and assessed its imputation performance (**Fig. 1b; Extended Data Fig. 1b**). Simulations across varying sequencing coverages revealed strong concordance for non-reference alleles, with accuracy increasing at higher coverage. At 10% coverage, agreement exceeded 60% across all populations and minor allele frequency (MAF) bins. Under these conditions, the mean dosage R^2 (DR^2) was 0.47, with 34.1% of the 17,109,424 imputed sites exceeding $DR^2 > 0.8$; applying a $DR^2 \geq 0.8$ threshold retained ~30% of the sites as high quality. The cumulative DR^2 distributions were consistent across populations, and the Pearson correlations between the mean dosages before and after imputation remained high ($r = 0.99$) at all coverages. Collectively, these results demonstrate that the haplotype imputation reference panel enables accurate and consistent imputation across diverse *H. pylori* populations, even at low sequencing depths. The panel is publicly available through HPgnomAD (<https://www.hpgnomad.top/>), where users can upload local data or download reference panels for imputation.

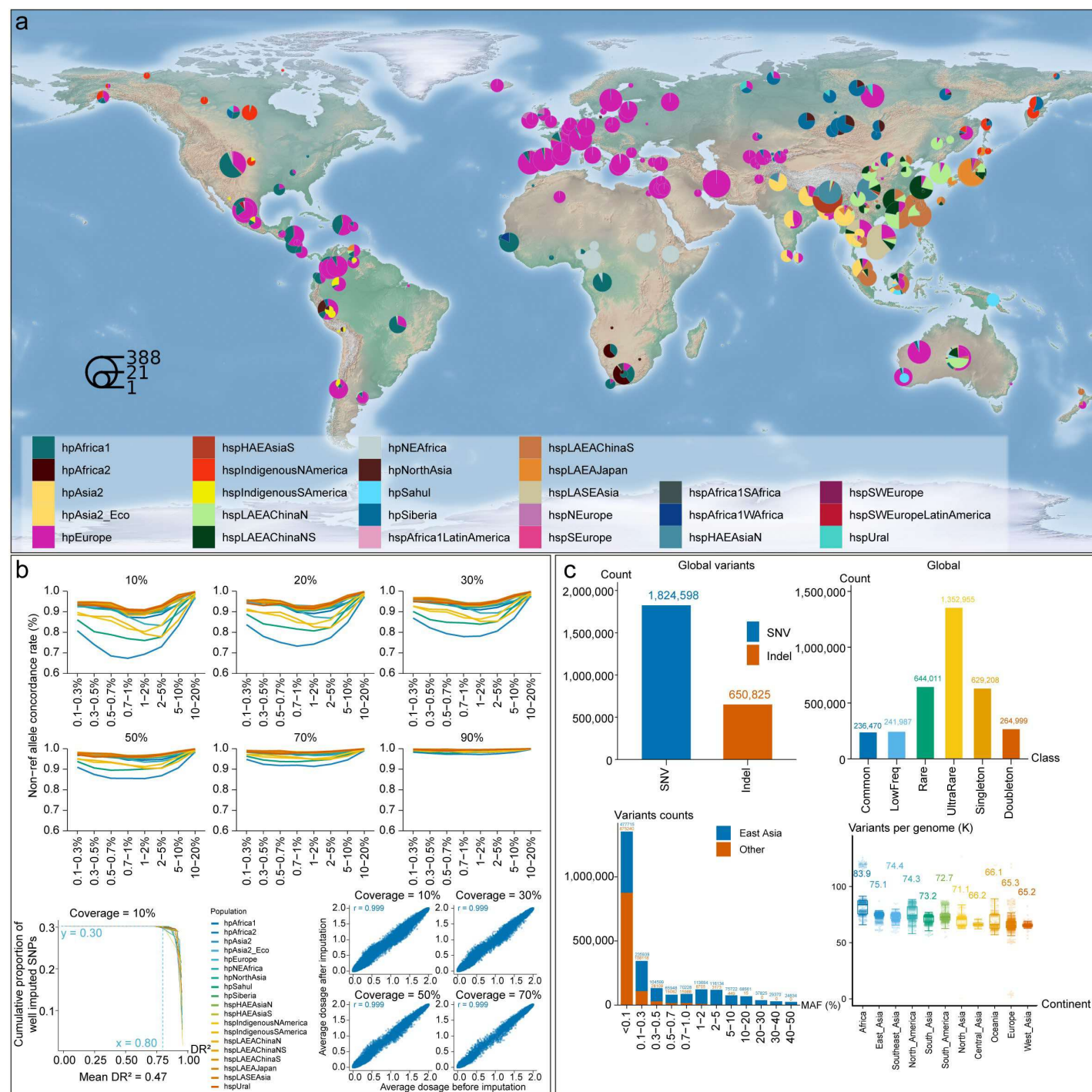
Using 7,544 high-quality *H. pylori* genomes, we characterized the distribution of SNPs and InDels. Overall, we found 1,824,598 SNPs and 650,825 InDels (**Fig. 1c**). These variants include 236,470 common, 241,987 low-frequency, 644,011 rare, and 1,352,955 ultra-rare sites, the latter comprising 629,208 singletons and 264,999 doubletons. Geographical analysis revealed that African strains have the greatest number of variants per genome (average: $\sim 83.9 \times 10^3$), reflecting a long evolutionary history and ancient origin. East Asian strains ranked second (average: $\sim 75.1 \times 10^3$), indicating unexpectedly high genetic diversity in this region. Among the 1,657 East Asian genomes, we identified 161,974 common, 222,578 low-frequency, 600,358 rare, and 427,723 ultra-rare variants (**Extended Data Fig. 1c**). The East Asian samples included nearly all the high-frequency sites ($MAF \geq 20\%$), whereas the $MAF < 0.1\%$ category alone contained 1,352,955 sites, which was the most considerable portion of the variants (**Fig. 1c**). Variant counts decrease sharply as the MAF increases, leaving only 24,834 sites in the 40~50% range. Low-frequency variants are likely recent in origin, influenced by region-specific selective pressures and demographic history, and may be associated with distinct clinical pathogenicity; thus, they merit detailed study. In East Asia, 1,518 of the 1,657 strains (91.61%) belong to the hpEAsia population (**Extended Data Fig. 1c**). However, this broad classification does not capture the complex genetic structure of these strains, highlighting the need for more refined classification.

Fine-scale genetic substructure and complex genomic diversity of hpEAsia strains

To explore the global genetic landscape of *H. pylori* strains, we first conducted principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) using the core genomes of 7,544 strains (**Fig. 2a and Extended Data Fig. 2**). The

129
130
131
132
133
134
135
136

overall patterns of *H. pylori* strains worldwide aligned with those recently described via HpGP, but more complex patterns appeared in East Asian strains because of the large sample size. We observed that although most East Asian *H. pylori* strains formed a distinct cluster, a subset showed a continuous gradient along PC1 that overlapped with some European strains (**Fig. 2a**). This pattern indicates historical mixing across Eurasia. When PCA points were colored based on genetic classification rather than geographic origin (**Extended Data Fig. 2**), hpEAsia strains again formed a separate cluster, distinct from other lineages along PC1. Conversely, hpAsia2, which is mainly found in mainland Southeast Asia and South Asia, remained genetically separate along PC2 despite being geographically close to East Asian strains, suggesting an independent evolutionary path.



137
138
139
140
141
142
143
144
145
146
147

Fig. 1 | Sample distribution in the HPgnomAD database, performance evaluation of haplotype imputation reference panel, and variant discovery. **a**, Geographical distribution of the samples used in this study. The colors indicate strain classifications, and the marker sizes correspond to the sample sizes. **b**, Reference panel performance evaluation. Upper panel: Genotype concordance by minor allele frequency (MAF) bin for each population, simulated at sequencing coverages of 10%, 20%, 30%, 50%, 70% and 90%. Lower left: cumulative distribution function of dosage R^2 (DR^2) at 10% coverage; the x-axis shows the DR^2 threshold, and the y-axis shows the fraction of total imputed SNPs with DR^2 . Lower right panel: Mean dosage before versus after imputation across coverage levels; Pearson correlation coefficients (r) are indicated. **c**, Variant discovery in the HPgnomAD database. Variants are classified by common (MAF $\geq 5\%$); low-frequency ($1\% \leq \text{MAF} < 5\%$); rare ($0.1\% \leq \text{MAF} < 1\%$); ultra-rare (MAF $< 0.1\%$); doubletons (variants observed in two samples); and singletons (variants observed in a single sample).

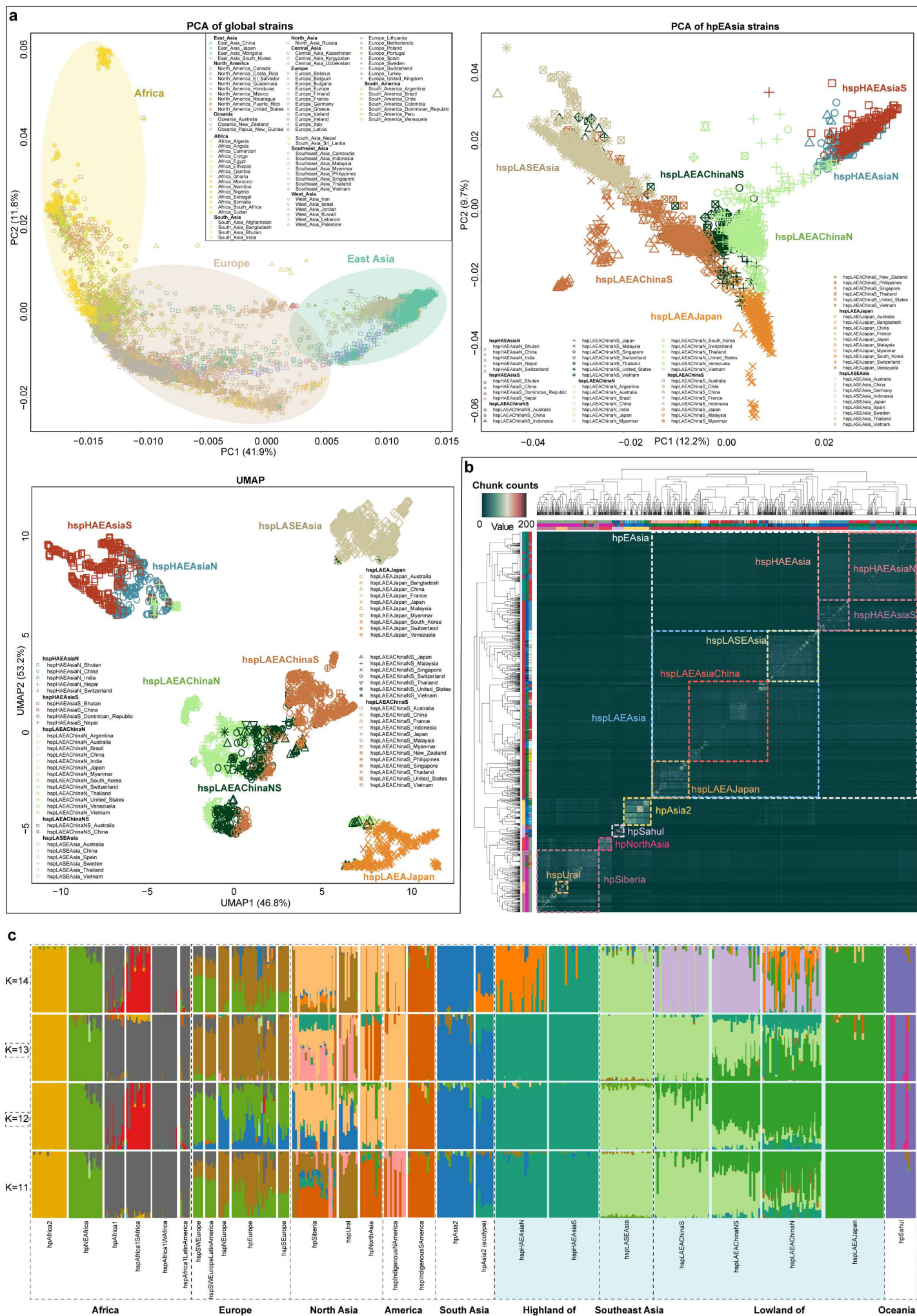


Fig. 2 | Genetic structure of global *H. pylori* strains. **a**, PCA of global and hpEAsia strains and UMAP of hpEAsia strains. For the global PCA, points are colored according to the sampling continent and shaped by the sampling country. For both the hpEAsia PCA and UMAP, the points are colored according to genetic background (see **Extended Data Fig. 2**). **b**, FineSTRUCTURE analysis of *H. pylori* strains (see **Extended Data Fig. 4** for more details). **c**, ADMIXTURE analysis showing the genetic structure of strains at $K = 11-14$ (see **Extended Data Table 2** for more details). The optimal K values, as determined by cross-validation, were 12 and 13. Shaded areas highlight hpEAsia strains.

Phylogenetic reconstruction revealed two distinct clades within hpAsia2. Genome-wide association testing, pairwise F_{ST} estimates, and analyses of genomic architecture collectively indicated that this clade represents an ecologically distinct population, hereafter called the hpAsia2 ecotype (**Extended Data Fig. 3**). For the hpEAsia strains, we used ordinary least squares regression on the strains from the East Asia subset to examine the relationships between geographic variables and the first three principal components (PCs) (PC1, PC2, and PC3). Each component was standardized through z-score transformation. Altitude was strongly positively correlated with PC1 ($R = 0.502$, $p < 0.001$), indicating clear differentiation between high- and low-altitude populations. Latitude was the most significant predictor of PC2 ($R = 0.178$, $p < 0.001$), suggesting that PC2 mainly captures north-south divergence (**Extended Data Table 2**). These results demonstrate that the genomic diversity and structure of East Asian *H. pylori* are closely influenced by geographic factors, especially altitude and latitude, emphasizing the key role of environmental variables, particularly altitude, in shaping population structure.

Furthermore, to characterize the fine-scale population structure of the hpEAsia group in our HPgnomAD dataset, we used fineSTRUCTURE to compute the coancestry matrix (**Fig. 2b** and **Extended Data Fig. 4**). The analysis revealed a complex demographic landscape among globally spread *H. pylori* lineages and significant diversity within the previously defined hpEAsia subpopulation. Following established naming conventions, we classified clusters by altitude (H/L) and latitude (N/S) within geographic regions. At the primary level, hpEAsia is split into hspHAEAsia (high-altitude) and hspLAEAsia (low-altitude) lineages. hspHAEAsia is subdivided into hspHAEAsiaN (northern Qinghai-Tibet Plateau) and hspHAEAsiaS (southern Himalayas, notably Bhutan), whereas hspLAEAsia is split into hspLASEAsia (Vietnam, Thailand), hspLAEACHina, and hspLAEAJapan. hspLAEACHina further separated into hspLAEACHinaS (southern China), hspLAEACHinaN (northern China), and a hybrid central cluster (hspLAEACHinaNS) enriched in admixed strains from the north-south transition zone.

To better understand the ancestral makeup of the hpEAsia subclusters, we performed a model-based ADMIXTURE analysis with a comprehensive global strain panel to reduce bias. Cross-validation revealed that $K = 13$ was the best fit for capturing the primary continental ancestry components that shape the admixture patterns. We observed meaningful genetic differences between East Asian *H. pylori* genomes and those from nearby regions such as North Asia, South Asia, and the Americas. In the most suitable model, three major components shaped hpEAsia diversity (**Fig. 2c**): a green component enriched in highland Lhasa and Bhutan, a light green component common in Southeast Asia, and a deep green component dominant in Japan, mirroring fineSTRUCTURE clusters. Chinese lineages presented varying admixture from all three lineages, with hspLAEACHinaNS exhibiting an intermediate profile between hspLAEACHinaS and hspLAEACHinaN. Increasing K to 14 separated hspHAEAsiaN and hspHAEAsiaS into distinct components, indicating unique genetic profiles and admixture histories. The concordance among ADMIXTURE, fineSTRUCTURE, and PCA underscores the presence of fine-scale, geographically structured genetic signatures within hpEAsia.

Deep evolutionary history of *H. pylori* indicates a temporal link between strain divergence and human migration

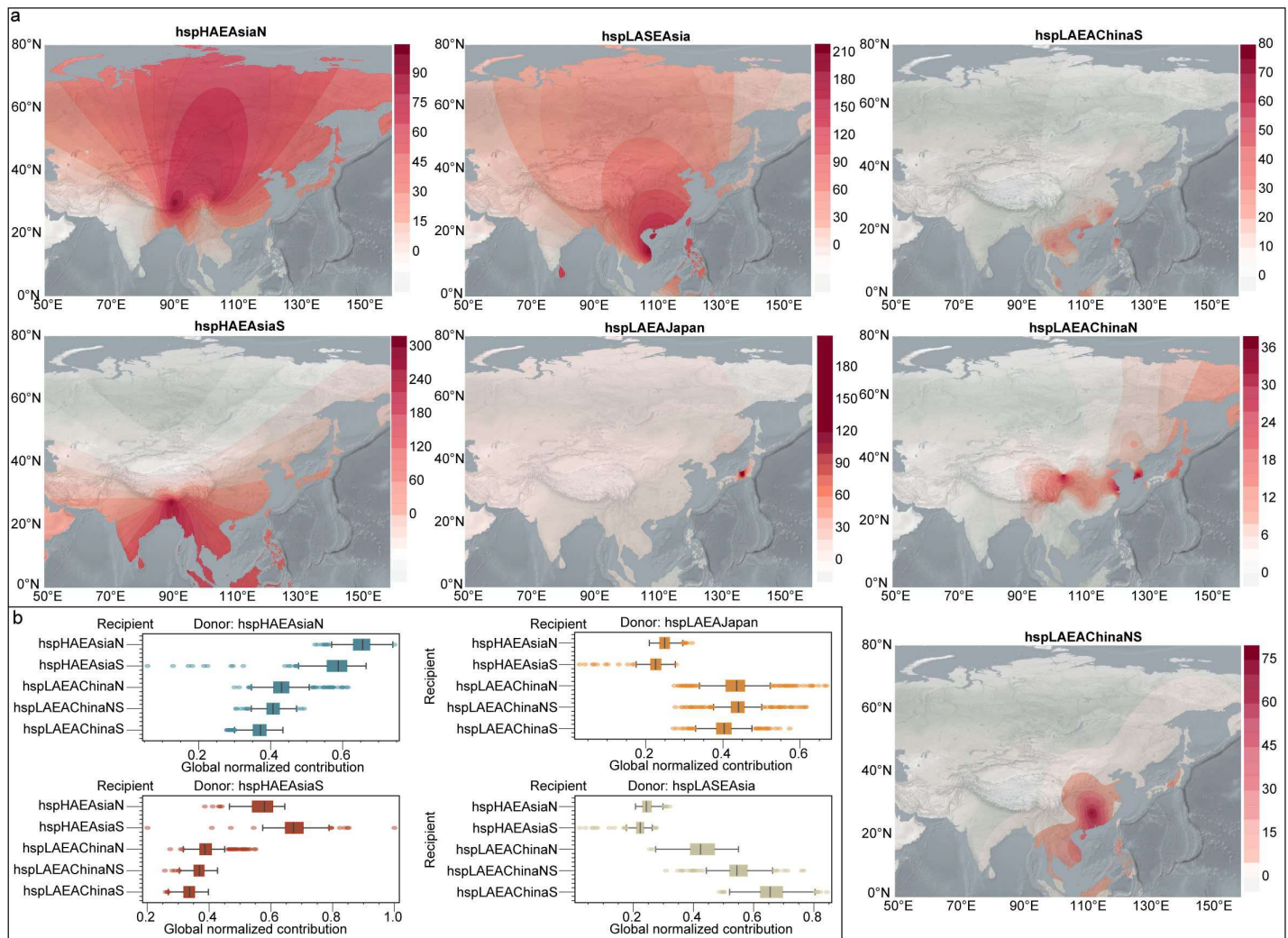
Previous genomic studies have reconstructed complex human migration patterns, including out-of-Africa dispersals and population splits during the late Paleolithic and Neolithic periods, via Multiple Sequentially Markovian coalescent models and BEAST-based Y-chromosome phylogenies^{21,23,33,63}. These studies also explored fine-scale human population structure and complex migration events. However, the early demographic history of *H. pylori*, such as its origin, initial expansion, bottlenecks, subsequent dispersal, and correlation with human genetic diversity patterns, remains poorly understood because of the underrepresentation of key lineages. To address this, we analyzed the spatial distributions of the major clusters and sublineages identified in our dataset (**Fig. 3a** and **Extended Data Fig. 5a**). Kriging-based contour maps displayed transparent spatial gradients for each strain. The two highland strains, hspHAEAsiaN and hspHAEAsiaS, had the highest predicted frequencies in the northern and southern Himalayan regions, respectively, which is consistent with their distinct human genetic origins and population structures⁶⁴. Notably, hspHAEAsiaN also presented elevated frequencies on the northeastern Qinghai-Tibet Plateau and in the upper Yellow River valley, indicating possible ancestral origins and migration routes^{65,66}. In contrast, hspLASEAsia and hspLAEAJapan reached peak frequencies in Southeast Asia and Japan, respectively.

Within the hspLAEACHina lineage, the northern sublineages were distributed mainly across the Yellow River Basin, northeastern China, and the Korean Peninsula. The secondary concentration in the Tibetan-Yi Corridor suggests southward migration from northern to southwestern China. The southern lineage was primarily found in the Zhujiang River Basin, along China's southern coast, and in Taiwan, with additional occurrences on certain Southeast Asian islands, indicating further southward dispersal from southern China into Southeast Asia. In the Yangtze River Basin and Central China, we observed high levels of admixture between the northern and southern sublineages. Correlation analyses between frequency and latitude and longitude for these sublineages revealed that the frequency of hspLAEACHinaN was positively correlated with latitude ($R = 0.421$, $p = 0.029$), whereas the frequency of hspLAEACHinaS was negatively correlated with latitude ($R = -0.415$, $p = 0.044$); neither lineage had a significant correlation with longitude (**Extended Data Fig. 5b**). These findings further support distinct genetic differences within the hspLAEACHina lineage. These patterns may reflect human demographic expansions linked to ancient millet farmers in northern China and rice cultivators in southern China^{65,67}, as well as geography-related population substructures inferred from Y-chromosomes and ancient DNA^{21-23,38}.

To further explore the genomic origins of the three Chinese lowland sublineages, we used ChromoPainter to infer haplotype sharing with geographically neighboring donor populations (**Fig. 3b**; **Extended Data Fig. 6**). We observed a north-south gradient in genetic input: hspLAEACHinaN was enriched for highland-derived haplotypes, whereas hspLAEACHinaS drew mainly from hspLASEAsia. The intermediate group, hspLAEACHinaNS, located in the transitional zone, showed mixed ancestry. Although primarily East Asian, these sublineages also included elements from non-East Asian sources; northern strains showed contributions

215
216
217
218
219

from hspLAEAJapan, hpNorthAsia, and hpSiberia, whereas southern strains shared features with Southeast Asian lineages^{22,67}. Additionally, these admixture patterns support hypotheses of southward expansion by southern Chinese farmers into Southeast Asia and the northward influence of millet-farming populations on the Qinghai-Xizang Plateau⁶⁵. Overall, widespread gene flow highlights the significant role of geography in shaping the evolutionary paths of *H. pylori* in East Asia.



220

Fig. 3 | Spatial structure of the hpEAsia lineage and ancestral profile of low-altitude isolates. a, Spatial distribution of the seven hpEAsia subtypes. Contour maps depict interpolated frequency surfaces generated with Kriging. **b**, Box-and-whisker plots showing ancestry contributions from selected donors to three recipient groups, including hspLAEACHinaS, hspLAEACHinaN, and the admixed hspLAEACHinaNS. Each panel represents one donor; the x-axis reports globally normalized contributions $\frac{x - \text{global}_{\min}}{\text{global}_{\max} - \text{global}_{\min}}$, and the y-axis lists recipient categories. Boxes indicate the interquartile range, whiskers span the full data range, and dots denote outliers. Only the donors with the highest contributions are shown; the complete set is provided in **Extended Data Fig. 6**.

225

To clarify the evolutionary relationships and phylogenetic history of globally representative *H. pylori* lineages, we performed maximum likelihood (ML) phylogenetic analysis to determine broad clustering patterns (**Fig. 4a and b**; **Extended Data Fig. 7a**). The clustering of non-Asian lineages matched recent reports by Tourrette *et al.*¹⁶. Notably, hpAsia2 diverged earlier than other Asian strains did and formed a distinct phylogenetic branch. Most other Asian strains were grouped into a large clade that shared a most recent common ancestor (MRCA) with hpSahul. The complex phylogenetic structure of strains from Siberia and Indigenous American populations, as previously noted, was beyond the scope of this study. Our focus was on East Asian strains, confirming that their clustering reflected geography-related population structure. Three main features emerged: first, multiple sublineages formed separate clades with unique genetic profiles; second, several lineages diverged nearly simultaneously, indicating rapid radiation; and third, despite overall genetic differences, some clades, particularly within hspLAEACHina, included other subgroups. This suggests extensive demographic interactions in East China, with frequent recombination leading to a mosaic evolutionary pattern in certain genomic regions. These findings were supported by an admixture graph generated with TreeMix (**Extended Data Fig. 7b**). The TreeMix analysis of Asian strains revealed that a model with five migration edges ($m = 5$) explained most of the variance, with additional edges offering minimal improvements ($< 0.05\%$). The topology closely matched the ML tree, with no substantial differences in population classification into major clades. Notably, hpEAsia populations presented short drift branch lengths, indicating limited genetic drift since divergence; this may reflect recent splits, large effective populations, or ongoing gene flow. Migration from the hpAsia2 ecotype into hspHAEAsia was also observed, implying historical genetic exchange between South Asia and the East Asian highlands.

245

To estimate divergence times, we reconstructed phylogenies via both ML and Bayesian inference (**Fig. 4c**; **Extended Data Fig. 8a**). The earliest strains in East Asia seem to be ancient, as both hpAsia2 and hpSahul are found in this region. Bayesian analysis estimated the MRCA of hpAsia2 and hpSahul to date back to 31.8 thousand years ago (kya) (95% highest posterior density [HPD]:

246
247

248
249
250
251
252
253
254
255
256

27.1–39.4 kya). Considering their current geographic distributions, these findings support a "southern dispersal" route for early human migration from Africa to the coastal regions of South and Southeast Asia^{68–70}. The MRCA of all hpEAsia strains was dated to approximately 23.6 kya (95% HPD: 21.9–29.4 kya), with the split between hspHAEAsia and hspLAEAsia estimated at 18.8 kya (95% HPD: 16.4–22.2 kya). The ML-based estimates were similar at 29.3 kya and 25.3 kya (**Extended Data Fig. 8b**). These results suggest at least two waves of human migration into East Asia, as shown in the bacterial phylogeny. Notably, the divergence between high- and low-altitude strains had already occurred by the Upper Paleolithic. Historical population dynamics revealed that the effective population size of global *H. pylori* increased in two phases: initially from the late Pleistocene to just before the Last Glacial Maximum (LGM), and subsequently from the LGM to just before the Younger Dryas (**Extended Data Fig. 9**). These expansions imply that climate-driven changes likely influenced human dietary habits and lifestyle shifts, which were later reflected in the *H. pylori* genome.

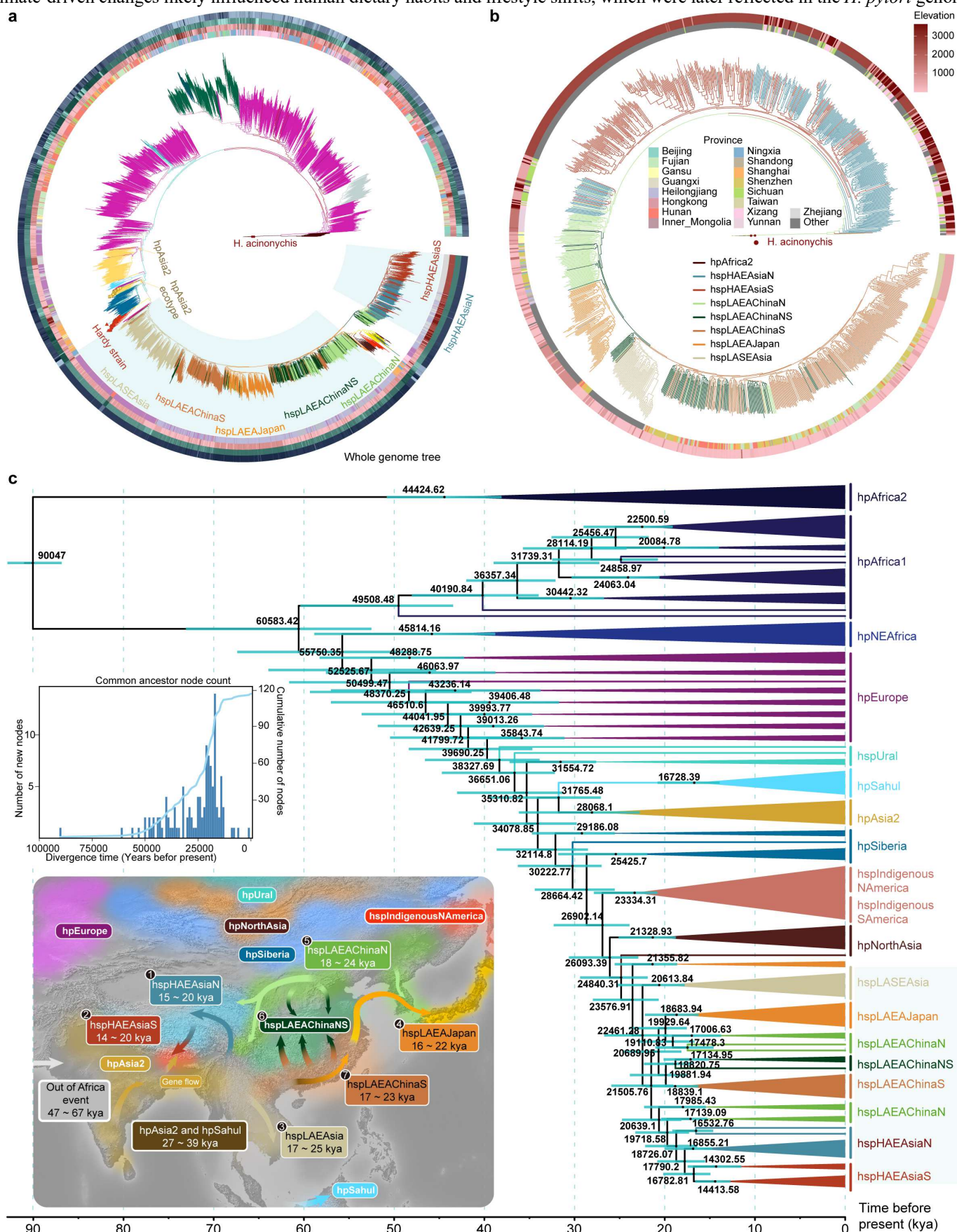


Fig. 4 | Phylogenetic analysis of *H. pylori* strains. **a**, Maximum-likelihood phylogenetic tree based on whole-genome sequences of the strains analyzed in this study. Branches are colored according to population. *H. acinonychis* serves as the outgroup, as indicated by a red square at the tip of its branch. Hardy strains are marked by red triangle dots at the branch tips, and the hpAsia2 ecotype is marked by a yellow square at the tip of its branch. The concentric circles surrounding each tree represent, from innermost to outermost, the sampling location's classification by continent, altitude, latitude, and longitude (for details, see **Extended Data Fig. 7a**). **b**, Phylogenetic tree of all hpEAsia strains based on whole-genome sequences. **c**, Time-scaled global phylogeny inferred by BEAST. Individual strains are colored according to subpopulation. At each node, light blue bars indicate the 95% HPD interval, and the mean node age is shown as a numerical label. The upper-left inset depicts the temporal distribution and demographic trajectory of ancestral nodes over successive intervals; the lower-left inset shows inferred migration routes and the timing of *H. pylori* dispersal events.

Biological adaptive signatures under varying geographic and environmental factors

Having established a strong hierarchical division and detailed population substructure of hpEAsia *H. pylori* populations, we aimed to identify loci and genes with distinctive allele frequency patterns linked to environmental and dietary differences between highland Tibetan and lowland East Asian hosts, as well as adaptive divergence between northern and southern East Asian lineages. Since PC1 and PC2 reflect altitude and latitude differences (**Extended Data Table 3**), we first compared the hspHAEAsia and hspLAEAsia lineages (**Fig. 5a**). This comparison revealed 256 SNPs with $F_{ST} > 0.6$, including 60 SNPs that also surpassed the GWAS significance threshold ($-\log_{10}p > 5$) (**Fig. 5b**; **Extended Data Table 4**). The strongest signals overlapped with neuraminylactose-binding, hemagglutinin-like adhesin, and outer membrane protein loci *alpA/hopC*. Additional notable loci with high scores included *frpB4*, *faaA2* (a *vacA*-like cytotoxin), *lysE*, *hopE*, and *hofF*. These genes are associated mainly with surface adhesion and iron acquisition, indicating functional adaptations to high-altitude environments. To investigate latitude-related differentiation, we performed a similar analysis comparing the hspLAEAsiaChinaS and hspLAEAsiaChinaN groups. This analysis revealed that 11 SNPs with $F_{ST} > 0.6$ and 65 SNPs reached GWAS significance ($-\log_{10}p > 5$) (**Fig. 5b**; **Extended Data Table 4**). Significant signals include the ribosome recycling factor, the *cag*-pathogenicity-island component *cagM*, and the outer membrane protein *hopC*. These findings suggest that immune modulation and cell-surface variation may play roles in the north–south divergence of *H. pylori* across East Asia.

To evaluate whether the observed patterns of varied genomic diversity indicate adaptive evolution, we calculated dN/dS ratios for both the combined coding sequence alignment (**Extended Data Fig. 10**) and individual CDSs (**Fig. 5c**). Among all lineages, hpEAsia presented a higher average dN/dS ratio than strains from Europe and the Americas did, with hspLAEAsiaJapan exhibiting the highest overall value. Within hspHAEAsia, increased dN/dS ratios were observed for *alpA/hopC*, the putative membrane protein-encoding gene *HP0622*, and *rpiB* (ribose-5-phosphate isomerase), whereas lower ratios were detected for genes encoding the 30S ribosomal protein S21, the DNA methyltransferase M. HpyAVIII, and the O-antigen polymerase (**Fig. 5c**). When the data were stratified by latitude, distinct signals of selection emerged. Genes involved in disulfide bond formation, including *flaA* (flagellin), phosphate acetyltransferase genes, and *hopC*, were targeted differently, indicating divergent selective pressures between northern and southern East Asian strains. To examine the functional basis of these adaptations, we annotated 1,577 coding sequences to identify pathways distinguishing high-altitude from low-altitude strains, as well as northern from southern populations. Despite the use of multiple annotation pipelines, 558 genes lacked Gene Ontology (GO) terms, and 446 lacked Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations, emphasizing the large fraction of *H. pylori* genes that remain uncharacterized (**Extended Data Fig. 11a**). We then overlapped genes with SNPs that passed both the GWAS ($-\log_{10}p > 5$) and the $F_{ST} > 0.6$ thresholds with those showing $|\text{Cliff's } \Delta| \geq 0.35$ in the dN/dS analysis to define a set of candidate targets. All other annotated genes served as the background for subsequent enrichment analyses.

Among highland–lowland high-differentiation adaptive loci (**Extended Data Fig. 11b**), no GO term or KEGG pathway remained significant after applying Benjamini–Hochberg correction ($FDR < 0.05$), indicating limited statistical power. However, several GO categories showed nominal enrichment (uncorrected $p < 0.01$). Among the top ten GO terms, eight represented biological processes, and two represented molecular functions. The enriched biological processes included siderophore transport ($p = 1.9 \times 10^{-4}$), iron coordination-entity transport ($p = 6.9 \times 10^{-4}$), and iron-ion transport ($p = 1.0 \times 10^{-3}$), along with nitrogen metabolism, specifically glutamine biosynthetic processes and nitrogen utilization ($p = 4.9 \times 10^{-3}$). In terms of molecular function, nominal enrichment was observed for energy-transducer activity and phosphomannomutase activity. Although none of these enrichments were significant after multiple testing correction, they aligned with recurrent signals of metal acquisition and surface adhesion identified through GWAS, F_{ST} , and dN/dS analyses. In the KEGG database, no pathways achieved significance after FDR correction. Nevertheless, pathways involved in carbon and nitrogen metabolism, two-component signaling systems, and *H. pylori* epithelial cell signaling presented the highest fold-enrichment values, albeit nominal, converging on processes related to metal uptake, nitrogen metabolism, and adhesion.

In the analysis of loci with significant north–south differentiation (**Extended Data Fig. 11c**), GO enrichment revealed no terms passing the $FDR < 0.05$ threshold. However, several categories were nominally enriched (uncorrected $p < 0.05$), mainly related to membrane transport and extracellular structure. These included efflux pump complexes, wide-pore channel activity, porin activity, and efflux transmembrane transporters, as well as flagellar structures such as bacterial-type flagella and cell projections, along with various transporter complexes. After trivial entries ($p = 1$ or fold enrichment = 0) were removed, three pathways were identified: a two-component system (fold enrichment = 25.7, $p = 1.5 \times 10^{-3}$), flagellar assembly (fold enrichment = 14.3, $p = 6.9 \times 10^{-2}$), and a bacterial secretion system (fold enrichment = 12.9, $p = 7.6 \times 10^{-2}$). Although these genes did not reach formal statistical significance, their focus on processes related to transmembrane transport (efflux/porin), motility (flagellum), and signal transduction (two-component systems) matches the genetic signals observed at *cagM*, *hopC*, and genes encoding flagellin and disulfide-bond enzymes. Overall, these results suggest that variation in cell-surface components and environmental sensing mechanisms may influence north–south genomic differences.

We expanded the environmental association analysis to include temperature, solar radiation, and water-vapor pressure, identifying numerous additional SNPs (**Extended Data Fig. 12a**). To reduce collinearity among the environmental variables, we performed PCA on the full environmental matrix (**Extended Data Fig. 12b**; **Extended Data Table 5**). The first two PCs explained 57.70% and 26.77% of the total variance, respectively. PC1 had positive loadings for temperature (+0.56), water-vapour pressure (+0.47), and solar radiation (+0.46) and negative loadings for latitude (−0.48) and altitude (−0.16), capturing a temperature–humidity–

insolation gradient. Higher PC1 scores corresponded to warmer, moister, and more irradiated environments. PC2 was driven primarily by altitude (+0.75) and solar radiation (+0.45), reflecting an altitude–radiation gradient, with higher scores indicating high-altitude, high-irradiance settings. We used PC1 and PC2 as composite environmental predictors in subsequent genome-wide scans (**Extended Data Fig. 12c**). PC1-based GWAS and F_{ST} analyses revealed clusters of significant SNPs in genes encoding the flagellar basal-body P-ring protein *flgI*, 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase, the outer-membrane protein *hopL*, type II citrate synthase *gltA*, and *alpA/hopC*. In contrast, PC2-based scans were dominated by signals in *firpB4*, the GTP-binding protein *obgE*, the vacA-like cytotoxin *faaA2*, a haemagglutinin-like protein, and *vacA*. These findings indicate that both geographic distance and localized environmental pressures jointly shape the population structure and adaptive landscape of *H. pylori* across East Asia. The resulting catalog of candidate SNPs and their genomic contexts provides a critical foundation for future functional studies aimed at elucidating microbial fitness traits, host–pathogen interactions, and disease potential in this globally relevant pathogen.

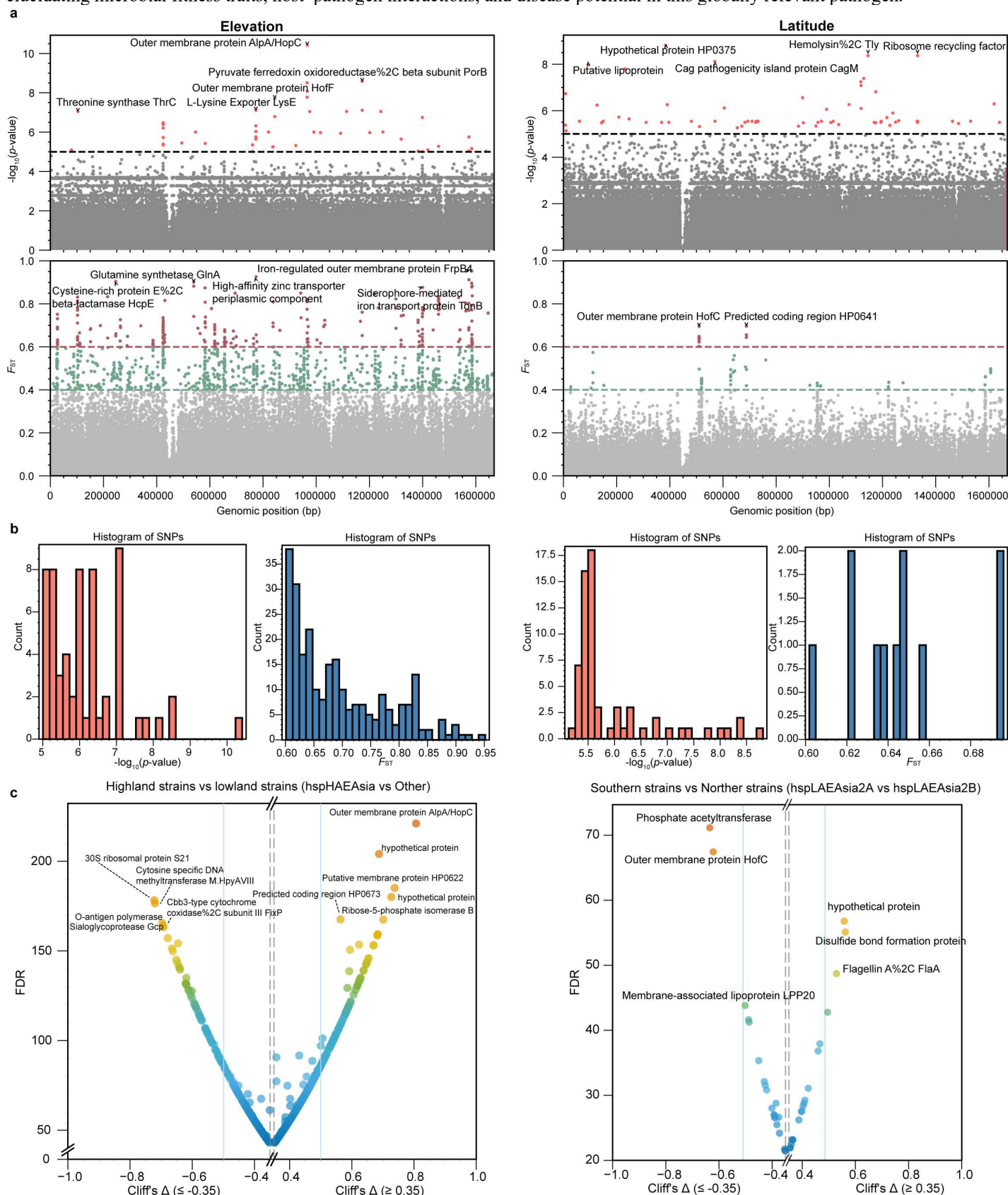
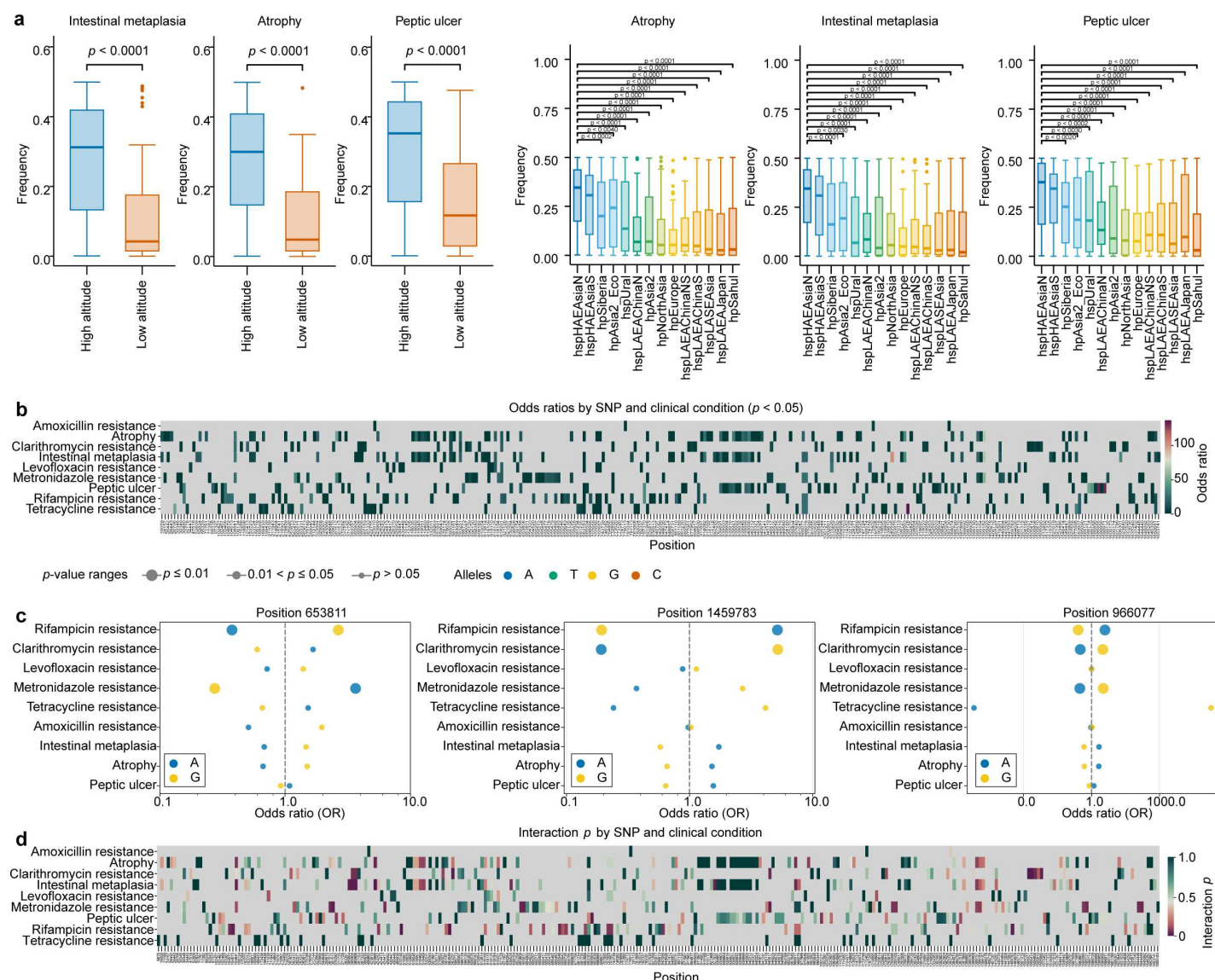


Fig. 5 | Altitude- and latitude-associated divergent loci in hpEAsia strains. a, GWAS and F_{ST} scans for variants correlated with

336
337
338
339
340
341
342

altitude and latitude. The GWAS significance is indicated by a horizontal threshold at $-\log_{10}p = 5$; sites exceeding this threshold are colored red, and those below it are gray. F_{ST} thresholds of 0.4 and 0.6 are marked, with SNPs colored gray (< 0.4), green ($0.4 \leq F_{ST} < 0.6$) or red (≥ 0.6). **b**, Counts of SNPs surpassing the respective GWAS or F_{ST} thresholds. The x-axis denotes $-\log_{10}p$ (GWAS) or F_{ST} , and the y-axis denotes the number of SNPs. **c**, dN/dS analysis per CDS. Only CDSs harboring altitude- or latitude-associated variants are shown; gene products were predicted with Prokka. The x-axis shows Cliff's Δ , and the y-axis shows the FDR-adjusted $-\log_{10}p$ (see **Methods**). The results with $|\text{Cliff's } \Delta| > 0.35$ are displayed; the ten CDSs with $|\text{Cliff's } \Delta| > 0.5$ are labeled.



343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365

Fig. 6 | Landscape of SNP variants associated with host clinical and antibiotic resistance phenotypes. **a**, Boxplots of allele frequency distributions of candidate loci across populations stratified by host clinical phenotype (peptic ulcer, intestinal metaplasia, gastric atrophy). Candidate loci are defined as the top 5% of SNPs from GWAS and F_{ST} analyses and intersect with the top 5% of loci from a high-versus-low-altitude differentiation scan; allele frequencies were computed for each locus within each population, and the statistical significance was tested using a two-tailed t-test. The distributions of loci associated with bacterial antibiotic resistance phenotypes are shown in **Extended Data Fig. 13c**. **b**, Odds ratios (ORs) from χ^2 or Fisher's exact tests ($p < 0.05$); color intensity represents OR magnitude on a custom linear scale, the color bar indicates OR values, and gray denotes non-significant associations. **c**, Bubble plot of ORs for pleiotropic SNPs. The x-axis shows ORs (log scale); the y-axis lists phenotypes. The bubble color indicates the allele (A or G), and the bubble size corresponds to the Bonferroni-adjusted p-value category ($p \leq 0.01$; $0.01 < p \leq 0.05$; $p > 0.05$). The dashed vertical line indicates OR = 1. ORs and p-values were obtained from logistic regression with Wald tests and Bonferroni correction; the remaining loci are shown in **Extended Data Fig. 14**. **d**, Heatmap of interaction p-values for SNP-phenotype associations; color intensity reflects p-value magnitude, and gray indicates non-significance.

Highly differentiated genomic variation shapes distinct clinical outcomes

We first demonstrated that *H. pylori* populations from high and low altitudes exhibit marked genetic stratification across multiple loci. To assess whether altitude-associated variants are linked to clinical traits, we analyzed 106 isolates from Lhasa and 133 isolates from Chengdu (**Extended Data Table 6**). Strain ancestry was inferred via ChromoPainter: genomes assigned to hspHAEAsia were classified as "high altitude", whereas those assigned to hspLAEAsia were classified as "low altitude". Clinical traits were categorized into two groups (**Extended Data Table 6**): (1) host gastric pathologies and (2) bacterial antibiotic resistance. For each trait, we performed both a GWAS and an F_{ST} scan (**Extended Data Fig. 13a, b**), selected loci in the top 5% of each analysis based on empirical thresholds, and then intersected these with the top 5% of loci from the high-versus-low-altitude differentiation scan. We then filtered out variants implicated in both altitude-related differentiation and clinical outcomes and compared their allele frequencies between

the high-altitude (hspHAEAsia) and low-altitude (hspLAEAsia) groups, as well as among individual strain populations (Fig. 6a; Extended Data Fig. 13c). A two-sample t-test confirmed highly significant differences in allele frequency between hspHAEAsia and hspLAEAsia ($p < 0.0001$), demonstrating that these loci were indeed differentiated by altitude. Moreover, both the hspHAEAsiaN and hspHAEAsiaS subpopulations presented higher allele frequencies than did all the other groups; overall, the hspLAEACHina population presented higher frequencies than hspLASEAsia and hspLAEJapan did. We then constructed contingency tables stratified by altitude and phenotype, applying Pearson's χ^2 test or Fisher's exact test as appropriate (see Materials and Methods). Odds ratios (ORs) were derived from the same contingency tables, and SNPs with $p < 0.05$ were considered significant. In total, 489 SNPs met this criterion; of these, 464 could be annotated by Prokka and mapped to 175 distinct genes (Fig. 6b; Extended Data Table 7). High ORs indicated strong positive associations between minor alleles and clinical traits. Notable examples included position 1,299,728 in the d-lactate dehydrogenase gene, associated with tetracycline resistance (OR = 136.5); position 1,586,955 in *frpB4*, associated with peptic ulcers (OR = 132); and position 1,240,595 in *gluP*, associated with intestinal metaplasia (OR = 101.7) (Fig. 6b). These SNPs, which vary in frequency between high- and low-altitude populations and correlate with distinct clinical outcomes, represent candidate risk alleles for future functional validation.



Fig. 7 | Web interface of HPgnomAD. **a**, Landing page providing an at-a-glance summary of the underlying dataset. **b**, Introductory module describing *H. pylori* biology and epidemiology. **c**, Query panel for entering a SNP position relative to strain 26695. **d**, Interactive results stratified by population and subpopulation, showing both absolute counts and allele frequencies. **e**, Interactive results stratified by country and continent, likewise displaying counts and frequencies. **f**, Dynamic map depicting the geographical distribution of allele counts at the queried locus. **g**, Dynamic map depicting the geographical distribution of allele frequencies at the same locus. **h**, Haplotype imputation reference panel and imputation function: users can upload a partial sequence file or download the reference panel for imputation.

Several loci were associated with more than one clinical phenotype. To investigate these pleiotropic effects, we applied logistic regression models to each SNP across multiple traits and identified 322 SNPs associated with at least two clinical phenotypes. Of these, 21 exhibited significant SNP–phenotype interaction effects (unadjusted $p < 0.05$), and following Bonferroni correction, 15 remained significant (Bonferroni-adjusted $p < 0.05$) (Extended Data Table 8; Extended Data Fig. 14), 14 of which mapped to nine distinct gene regions. Notably, SNP 653,811 in *faaA2* (putative VacA-like toxin) was significantly associated with resistance to both rifampicin and metronidazole ($p < 0.05$ for each) (Fig. 6c). Similarly, the A allele at position 1,459,783 conferred increased rifampicin resistance (OR = 5.18, $p < 0.05$), whereas the G allele at the same site was linked to clarithromycin resistance (OR = 5.22, $p < 0.05$). SNP 966,077 in *alpa/hopC* correlated with resistance to rifampicin, clarithromycin, and metronidazole ($p < 0.05$ for all). These findings identify a set of pleiotropic variants whose mechanistic underpinnings warrant further investigation.

To assess whether the clinical impact of specific variants is influenced by the genomic background shaped by altitude, we implemented genotype–altitude interaction models for each SNP, and we retained 20 SNPs with interactions $p < 0.05$, 19 of which mapped to 12 distinct gene regions (Fig. 6d; Extended Data Table 9). These SNPs indicated that the clinical consequences of identical mutations varied between high- and low-altitude strains. For rifampicin resistance, mutations at position 510,145 (G>T) in *hofC* (OR = 14.3) and at position 583,049 (G>T) in *cagA* (OR = 11.7) significantly increased rifampicin resistance in high-altitude isolates. Similarly, a variant at position 385,904 (A>G) in *dsbC* was strongly associated with an elevated risk of intestinal metaplasia in high-altitude hosts (OR = 11.7). Together, these altitude-dependent variants highlight promising candidates for functional validation

and offer potential biomarkers for precision diagnostics, individualized treatment, and region-specific clinical management strategies.

Integrated database and query platform

To address the fragmentation and inconsistent updates among existing resources, we developed a unified, web-based database and query platform that enables rapid access to genome-wide SNP variation data in *H. pylori* (Fig. 7a-g). Users can input a genomic coordinate corresponding to an *H. pylori* SNP site, prompting the system to retrieve variation data across 7,544 whole-genome sequences. The platform then computes the count and frequency of each allele, accounting for the possibility of multiple alleles at a single locus, under six classification schemes: Subpopulation: 26 groups defined in this study, including six newly reclassified lineages (hspLASEAsia, hspLAEAShinaS, hspLAEAShinaNS, hspLAEAShinaN, hspHAEAShinaN, and hspHAEAShinaS) derived from the former hpEAsia, one hpAsia2 ecotype subset, and 19 lineages adopted from previous studies. Geographical bubble map, where bubble size reflects allele count and color represents allele type at each location. An animated density map illustrates allele frequency (allele count divided by total count at each site) across geographic coordinates. There is also information about the main population, country, and continent. By providing immediate visualization of the global allele distribution at any locus, the platform enables researchers to identify lineage- or region-specific SNPs, thereby accelerating the generation of hypotheses, experimental design, and development of targeted public health strategies. The platform also incorporates an *H. pylori* reference panel module for genotype imputation (Fig. 7h). Users can upload draft assemblies, after which the cloud server imputes missing genotypes and returns a complete whole-genome file. To support local analyses and reduce the server load, the reference panel itself is available for download and offline imputation.

Discussion

Fine-scale population substructure and deep demographic history of *H. pylori*

H. pylori colonizes the human stomach almost exclusively, with infection recognized as the principal etiological driver of atrophic gastritis and a major risk factor for peptic ulcer disease and gastric carcinoma. Genomic analysis of *H. pylori* offers a dual lens, illuminating human origins, migrations, and evolutionary history while clarifying host-pathogen co-evolution. These insights provide a robust foundation for disease prevention, precision therapy, and personalized medicine. Epidemiological data indicate that gastric inflammation severity and gastric cancer risk are markedly greater in Asia, particularly East Asia, than in Europe and North America, a disparity likely driven by more virulent *cagA* and *vacA* genotypes⁷¹⁻⁷³. This underscores the urgency of comprehensively characterizing East Asian *H. pylori* strains. Earlier genomic investigations, constrained by available technologies, relied heavily on MLST, pulsed-field gel electrophoresis, random amplified polymorphic DNA, and amplified fragment length polymorphism, with MLST as the dominant method. MLST, which targets seven housekeeping genes, enables the classification of seven modern and six ancestral *H. pylori* populations⁷⁴, but its limited genomic scope restricts resolution. The advent of high-throughput WGS supplanted MLST, with the first *H. pylori* genome sequenced by Tomb *et al.* in 1997⁹, ushering in fine-scale reconstructions of global population structure and migratory history^{15,16,29,30,51,58}. Despite East Asia's vast geography, demographic complexity, and host genetic diversity, most studies have continued to treat its *H. pylori* strains as a single hpEAsia (or hspEAsia) population^{15,16}. Although some studies have identified regional sublineages and proposed location-based nomenclatures^{29,30,75}, these efforts have been limited by regional sampling, incomplete temporal resolution, and the continued use of MLST-based definitions, which obscure whole-genome diversity and microvariation. The Qinghai-Tibet Plateau, a high-altitude, hypoxic region with a distinct human population history, remains a major gap, with scarce isolates hindering evolutionary reconstruction in this extreme environment. To address these gaps, we assembled the largest *H. pylori* genomic dataset to date, integrating 7,544 publicly available genomes. Using fineSTRUCTURE, PCA, and ADMIXTURE, we present the first genome-wide hierarchical subdivision of hpEAsia, revealing pronounced differentiation between high- and low-altitude communities and along a north-south latitudinal gradient.

Our expanded dataset uncovers a far more complex hpEAsia substructure than previously recognized. To avoid geographic nomenclature that obscures relationships, we adopt an alphanumeric scheme. At the broadest scale, hpEAsia segregates into two altitude-defined lineages: hspHAEAsia (high-altitude) and hspLAEAsia (low-altitude). hspHAEAsia is confined to the Qinghai-Xizang Plateau and adjacent regions, comprising hspHAEAsiaN (Tibet, China) and hspHAEAsiaS (Bhutan). hspLAEAsia includes three main sublineages—hspLASEAsia (predominant in Southeast Asia), hspLAEAShina (widespread in mainland China), and hspLAEAJapan (largely restricted to Japan). Within hspLAEAShina, ADMIXTURE, fineSTRUCTURE, and phylogeny revealed additional fine-scale substructures including hspLAEAShinaS, hspLAEAShinaN, and an admixed hspLAEAShinaNS. Regression analyses confirmed that both altitude and latitude significantly influence genomic variation, with altitude exerting a stronger effect. While the hspHAEAsia-hspLAEAsia split remains the deepest division, a north-south gradient persists within hspLAEAShina, with transitional populations reflecting continuous gene flow among mainland Chinese hosts. ChromoPainter analysis revealed that each hspLAEAShina subtype derives more genomic segments from nearby strains than from distant strains, mirroring historical human migration and bacterial exchange. In addition to these subdivisions, we identified additional genomic divergence and a novel ecotype within hpAsia2. Tourrette *et al.* defined an ecotype as a bacterial population that maintains a single recombining gene pool but exhibits speciation-level divergence in specific genomic regions¹⁶. In the study of hspIndigenousS America and hspIndigenousN America, they reported highly recombining yet strongly differentiated genomic segments, challenging classical species-formation models. Our findings demonstrate that *H. pylori* ecotypes are widespread, expanding the ecological and evolutionary framework for ecotype diversification.

TMRCA estimates from ML and Bayesian analyses date the split between hpAfrica2 and all other lineages to ~90.0 kya (95% HPD: 86.9–92.9 kya). hpNEAfrica and non-African strains diverged at 55.8 kya (95% HPD: 52.5–73.1 kya), which aligns with previous *H. pylori*⁷⁶, human Y-chromosome⁷⁷ and mtDNA⁷⁸ estimates. The earliest Asian lineage, hpAsia2, diverged from hpSahul ~31.8 kya (95% HPD: 27.1–39.4 kya), implying modern human expansion into South Asia and the Sahul by that time. hpEAsia's TMRCA is ~23.6 kya (95% HPD: 21.9–29.4 kya), with the hspHAEAsia-hspLAEAsia split at ~18.8 kya (95% HPD: 16.4–22.2 kya), firmly in the Paleolithic. Archaeological^{79,80} and genetic⁸¹⁻⁸³ evidence support high-altitude habitation on the Qinghai-Xizang Plateau by ~18.8 kya. Ancient DNA^{23,84} further documented a north-south East Asian split by ≥ 19 kya, matching the diversification of hspLAEAShina. Together, pathogen and host evidence converge on both altitudinal and latitudinal differentiation in East Asian human-*H. pylori* histories. Reconstructing *H. pylori* evolutionary relationships, however, requires careful treatment of three challenges:

pervasive recombination, extreme genetic diversity, and molecular clock uncertainty^{44,85,86}. First, *H. pylori*'s high recombination rate introduces more variation than de novo mutations do. While tools such as ClonalFrameML can remove recombinant segments for divergence dating, this often erases recent demographic signals, producing artificially flat BSPs⁸⁷. Second, *H. pylori*'s small yet highly diverse genome (~1.6 Mb) yields a nonlinear increase in segregating sites with larger sample sizes, creating computational bottlenecks for Bayesian algorithms. Third, mutation rate estimates remain incomplete, with microevolutionary rates exceeding macroevolutionary rates because of purifying selection^{44,88}. Robust molecular dating demands calibration against well-dated historical or geological events, with ancient DNA providing key opportunities⁸⁹, as in human mtDNA⁹⁰⁻⁹² and *Mycobacterium tuberculosis*⁹³ studies. Future work should focus on developing frameworks that integrate recombining and nonrecombining genome regions to infer evolutionary history and gene flow jointly, designing phylogenetic algorithms optimized for large-scale WGS analyses, and refining molecular-clock estimates through calibration with ancient *H. pylori* genomes.

Biological adaptive signatures and clinical significance

Previous studies reported a greater risk of *H. pylori* infection in Tibetan populations than in Han populations⁹⁴. An MLST-based analysis attributed this susceptibility to a Western-type (hpEurope) *cagA* allele in Tibetan isolates⁹⁵. In contrast, our whole-genome analyses revealed that plateau strains belong almost exclusively to hspHAEAsiaN and showed no clustering with hpEurope in either genome-wide or core-CDS phylogenies, underscoring the necessity of WGS for accurate strain classification. To pinpoint the genomic features driving high-altitude versus low-altitude differentiation, we performed a GWAS and calculated the F_{ST} , identifying numerous highly differentiated candidate loci likely contributing to increased virulence in plateau strains. dN/dS analysis highlighted a set of genes encoding hemagglutinin-like adhesin, the outer-membrane protein *alpA/hopC*, *frpB4*, and several iron-uptake system components. GO enrichment indicated nominal (uncorrected $p < 0.01$) overrepresentation of iron-transport terms, including siderophore transmembrane transport ($p = 1.9 \times 10^{-4}$), iron–ligand transmembrane transport ($p = 6.9 \times 10^{-4}$), and iron ion transport ($p = 1.0 \times 10^{-3}$), as well as nitrogen metabolism categories. Although these signals did not survive multiple testing correction, they align with the GWAS, F_{ST} , and dN/dS results, suggesting that plateau strains adapt to the low-iron gastric environment by optimizing iron acquisition and adhesion, enhancing colonization and pathogenicity.

North–south genetic differentiation was relatively modest. The most differentiated SNPs in these contrasts mapped to genes involved in cell-surface architecture and signal transduction, such as *cagM*, *hofC*, flagellin, and disulfide bond-forming enzymes, indicating a subtler latitudinal signature of adaptation. We caution against attributing all differentiation solely to altitude or latitude, as correlated environmental factors may also impose selective pressures. Environmental principal component analysis decomposed multiple climatic variables into two components, implicating composite gradients of temperature–humidity–radiation and altitude–radiation as major drivers of *H. pylori* genomic adaptation. These findings lay the groundwork for future functional validation of adaptive evolution in this pathogen. To assess potential clinical impacts, we compiled phenotypic data on antibiotic resistance (tetracycline, rifampin, metronidazole, clarithromycin, etc.) and gastric disease manifestations (peptic ulceration, intestinal metaplasia, etc.). GWAS and F_{ST} analyses identified SNPs linked to divergent clinical outcomes, addressing three questions: whether genomic variants distinguishing high- and low-altitude strains are associated with host pathology or drug resistance; whether individual SNPs show pleiotropic effects; and whether identical mutations produce distinct clinical consequences in plateau versus lowland contexts. Multiple candidate SNPs have emerged, with several conferring stronger pathogenic or resistance phenotypes in high-altitude isolates—evidence for genotype–environment interactions shaping clinical presentation. These variants offer promising biomarkers for precision diagnostics and individualized therapy, enabling targeted interventions across ecological contexts. Recognizing the critical role of WGS in resolving population structure and mapping pathogenic loci, we aggregated genome-wide variant calls from 7,544 isolates into an online query platform. This resource allows users to retrieve allele counts and frequency distributions for any SNP across strains and regions, which is supported by interactive visualizations. By integrating large-scale genomic data with population-genomic analyses, we provide new perspectives and tools for elucidating *H. pylori* evolution, environmental adaptation, and clinical diversity, advancing both basic understanding and precision public health strategies.

Materials and methods

Sequence collection and quality control

A total of 8,010 *H. pylori* sequences were retrieved from publicly available databases, including NCBI (<https://www.ncbi.nlm.nih.gov/>), EnteroBase (<https://enterobase.warwick.ac.uk/>), PubMLST (<https://pubmlst.org/organisms/helicobacter-pylori/>), and Datadryad (<https://datadryad.org/stash>). After quality control, 7,544 high-quality sequences (including two *H. acinonychis* sequences and 7,542 *H. pylori* sequences) were retained. We generated a new *H. pylori* genomes with detailed clinical phenotypes from 106 individuals in Lhasa, Tibet, and 89 individuals in Chengdu, Sichuan, using the Illumina MiSeq platform in this database. To ensure the integrity of downstream variant discovery and annotation, the reference genome was preprocessed by creating a genome index with Samtools and a sequence dictionary with Picard, providing the necessary metadata for variant analysis. Initial variant filtering employed GATK (v4.2.6.1) VariantFiltration with the following thresholds: $QUAL < 20$, $QD < 1$, $MQ < 30$, $FS > 100$, and $SOR > 10$ ⁹⁶. Variants flagged as low quality were subsequently removed using VCFtools (v0.1.16). To further maximize reliability, we applied stringent criteria, retaining only variants with a missing rate $\leq 5\%$, minimum read depth ≥ 3 , $MAF \geq 5\%$, average sequencing depth ≥ 20 , and $QUAL \geq 30$. Low-quality genomes were excluded on the basis of the following criteria: assembly fragmentation exceeding 500 contigs and coverage of the reference genome (*H. pylori* strain 26695, NC_000915.1) below 70% using BWA v.0.7.13⁹⁷, as described by Tournette *et al.*¹⁶. Genome annotation for each sequence was performed via Prokka v1.14.6 (option: --genus Helicobacter, --species pylori, --kingdom Bacteria, --evaluate 10^{-24})⁹⁸.

Core genome variant calling

The nucmer module in MUMmer v4 software was employed for SNP calling⁹⁹. Specifically, each genome sequence in the dataset was aligned against the reference genome of *H. pylori* (strain 26695, NC_000915.1). SNPs were identified via the show-snps and show-coords functions with the parameters (-ClrT and -rcIT). Delta files and coord files were converted into VCF format as described

previously¹⁶. BCFtools v.1.14 was used to merge files, summarize variant information, and extract specific variants^{100,101}. In a total of 7,544 *H. pylori* genomes, we identified 1,864,491 SNPs and 497,548 InDels. After a minimum allele frequency threshold of 1% was applied, 829,724 core SNPs and 271,723 InDels were retained. To simplify the alignment, we applied snp-sites v2.5.2 to remove monomorphic positions, retaining only polymorphic sites¹⁰². GNU parallel software was used to run multiple serial command line programs in parallel¹⁰³. The visualizations were generated via the Tidyplots package¹⁰⁴.

Haplotype imputation reference panel construction

For HPgnomAD, raw variants were first processed with BCFtools by removing SNPs located within 3 bp upstream or downstream of indels^{100,101}. Uncertain genotypes were set to missing, and all missing calls were filled with the reference allele to generate a standardized haploid haplotype reference. Non-polymorphic sites (allele count = 0) were then filtered out via BCFtools view. Each sample set was subsequently phased and self-imputed with Beagle 5.0¹⁰⁵. To evaluate imputation performance, we randomly selected 20 individuals per population as validation samples. Within these samples, genotypes were masked at different coverage levels to simulate low-information scenarios, and imputation was performed separately with each panel via Beagle 5.0. Imputation quality was assessed for each SNP by three complementary metrics: (1) dosage correlation squared (DR^2), as estimated by Beagle on the basis of the ratio of imputed dosage variance to the theoretical variance under known genotypes; (2) empirical dosage R^2 (R^2), calculated as the square of the Pearson correlation between imputed and true dosages across all validation samples; (3) the Pearson correlation (r) between the mean dosage before and after imputation; and (4) genotype concordance, defined as the proportion of best guess imputed genotypes that exactly match the true genotypes. By combining these metrics and stratifying by masking level, we quantified the panel's accuracy and robustness under low coverage or low-density conditions.

Strain assignment

According to previous studies^{12-14,16,17}, 285 strains were consistently assigned to one of 19 *H. pylori* populations/subpopulations. These 285 strains were previously examined and confirmed to represent their respective populations or subpopulations. To optimize computational efficiency, we randomly selected ten strains from each population, retaining all available strains when fewer than ten existed (**Extended Data Table 1**). We then assembled a dataset of 791 strains and applied fineSTRUCTURE¹⁰⁶ to validate both their individual assignments and population assignments. To dissect the genetic substructure of the hpEAsia lineage, we iteratively removed non-hpEAsia strains and repeated the fineSTRUCTURE analyses (**Extended Data Fig. 2**). The fineSTRUCTURE analysis was carried out for 200,000 Markov chain Monte Carlo iterations, including a burn-in of 50,000 iterations, based on core genome SNPs. After verifying the reassignment of each strain, these strains were used as a donor panel to "paint" each sample of remaining hpEAsia via ChromoPainter v2¹⁰⁶. Each genome was then assigned to a population on the basis that the population contributed the largest proportion of its painted genomic segments.

To objectively evaluate the difference in contributions between the hspLAEACHinaS and hspLAEACHinaN components derived from the ChromoPainter analysis, we constructed a differential index defined as follows: $\delta = \text{hspLAEACHinaS} - \text{hspLAEACHinaN}$. Statistical analysis of the δ values from 1443 samples yielded a standard deviation of $\sigma \approx 249$. To distinguish samples with minimal contribution differences, we adopted a relative thresholding approach by setting the threshold as $k \times \sigma$. In this study, we selected $k = 0.5$, corresponding to a threshold of approximately 125. Therefore, samples satisfying $|\delta| \leq 0.5 \times \sigma$ were considered to exhibit negligible differences between hspLAEACHinaS and hspLAEACHinaN contributions and were subsequently classified into a hybrid group, denoted hspLAEACHinaNS (**Extended Data Fig. 7**). Finally, we delineated 26 populations and subpopulations from 7,544 genomes (**Extended Data Table 1**). We subdivided hpEAsia into seven subtypes, including hspHAEAsiaN, hspHAEAsiaS, hspLAEACHinaN, hspLAEACHinaNS, hspLAEACHinaS, hspLAEAJapan and hspLASEAsia, and identified a distinct ecotype within hpAsia2, designated the hpAsia2 ecotype.

Principal component analysis

The merged VCF format file was processed via PLINK v1.9 to remove linkage disequilibrium according to the following criteria: window size of 50 base pairs, step size of 10 variants, and an r^2 threshold of 0.1¹⁰⁷. Visualization was performed via Python v3.12.2, which employs the seaborn and Matplotlib packages in Python¹⁰⁸⁻¹¹⁰.

Uniform Manifold Approximation and Projection (UMAP)

To investigate the genetic structure of hspHAEAsia and hspLAEAsia, we performed UMAP dimensionality reduction in Python v3.12.2 with the umap-learn package (v0.5.3). The parameters were set as follows: $n_neighbors = 10$, $min_dist = 0.1$, $n_components = 2$, $metric = \text{Euclidean}$, $init = \text{spectral}$, and $random_state = 42$ ¹¹¹.

Spatial geographic distribution

The spatial distribution of each strain across geographic Asia was visualized via Surfer[®] from Golden Software, LLC (www.goldensoftware.com). Contour maps depicting geographic frequency patterns were generated via the Kriging interpolation algorithm implemented in Surfer. Data derived from non-population-based samples were excluded from the analysis to ensure representativeness.

Genome structure comparison

To compare genome architecture, we employed dot plots to visualize substantial structural variations, such as inversions, gaps, repeats, and gene cluster rearrangements. Dot plots were generated with Gepard (v. 1.40) to compare genomic similarity and divergence between two distinct hpAsia2 lineages within the phylogeny¹¹². Specifically, we carried out pairwise comparisons of the hpAsia2 ecotype versus hpAsia2, hpAsia2 ecotype versus hpAsia2 ecotype, hpAsia2 versus hpAsia2, and hpAsia2 ecotype versus *H. cetorum* (NCBI accession CP003479.1). The alignments and plots were constructed with Gepard's internal DNA substitution matrix (edna.mat). A lower color threshold of 10 was chosen to suppress background noise and emphasize regions of interest. Both the window size and word length remained at their default settings (0 and 10, respectively).

Phylogenetic analysis

We conducted phylogenetic analyses of the whole genome and the core coding sequence (CDS) regions. The aligned sequences were then used to construct phylogenetic trees via VeryFastTree v4.0.4 with the -nt option¹¹³. The resulting tree files are visualized and refined via the ggtree and ggtreeExtra packages^{114,115}. For each tree, *H. acinonychis* served as an outgroup. To estimate strain divergence times, we applied two established methods^{15,30}, using a stratified random subset of 100 samples from the dataset to increase computational efficiency. In the first approach, we reconstructed a phylogeny via IQ-TREE v3 (options: -m MF, -T AUTO) and Model Finder under the Bayesian information criterion to select the GTR+G substitution model¹¹⁶. We then applied ClonalFrameML v1.12 to these whole-genome sequences under default settings to account for recombination¹¹⁷. Node ages were estimated in PAML v4.10.8 (model = 7, Mgene = 0, ncatG = 4)¹¹⁸ by assuming a uniform mutation rate of 1.38×10^{-5} substitutions per site per year⁴⁴. In the second approach, recombinant regions were first masked via ClonalFrameML v1.12. The filtered alignment was then analyzed in BEAST2 v2.5.0¹¹⁹, running 4×10^7 MCMC iterations under a GTR substitution model with a relaxed lognormal clock and the same mutation rate of 1.38×10^{-5} per site per year.

Admixture graphs

ADMIXTURE analysis was conducted via ADMIXTURE software¹²⁰. To optimize computational efficiency, we first subsampled global strains by randomly selecting 20 isolates from each non-hpEAsia population or subpopulation (retaining all available isolates when fewer than 20 were present) and 50 isolates per hpEAsia population. We then evaluated K values from 10 to 20 via eight computational threads, ten-fold cross-validation and a fixed random seed (12345) to ensure reproducibility. Admixture graphs among the different populations were constructed via TreeMix v.1.12¹²¹. The analysis was performed with migration edges ranging from 0 to 15, with 10 replicates for each specified number of edges. The population hpAfrica2 was designated as the outgroup. The optimal number of migration edges was determined as the smallest value that accounted for 99.8% of the explained variance.

The calculation of nonsynonymous and synonymous substitutions

Nonsynonymous substitutions refer to nucleotide changes that result in amino acid changes, whereas synonymous substitutions are those that do not alter the amino acid sequence. Both substitutions and their ratios were determined through pairwise comparisons of each strain's genome coding sequences against the hpAfrica2 strain via PAML v4.9¹¹⁸. The calculations were performed via the Nei-Gojobori (1986) method¹²² and the Yang-Nielsen (2000) method¹²³.

To pinpoint CDS regions under natural selection in *H. pylori*, we computed dN/dS for all 1,577 CDS regions annotated by the reference strain 26695 (NC_000915.1) against the hpAfrica2 strain via the Nei-Gojobori framework. Differences in selective pressure were then assessed via two complementary strategies: (1) the strain of interest was treated as one group, and all remaining strains constituted the second group; (2) two arbitrarily chosen strain sets were compared directly. Statistical tests were restricted to CDS regions represented by ≥ 5 observations in both groups. The distributions of dN/dS were contrasted with a two-sided Mann-Whitney U test, yielding U statistics and raw P values. The effect size was expressed as Cliff's Δ , which was calculated as $\Delta = \frac{2U}{n_1 n_2} - 1$, where n_1 and n_2 denote the sample sizes of the two groups. Multiple testing correction was performed with the Benjamini-Hochberg procedure. CDS regions meeting $FDR < 0.05$ and $|\Delta| \geq 0.33$ were deemed to exhibit significant selective differences.

GWAS and F_{ST} analysis

The environmental factors, external conditions influencing organismal development, behavior, and survival, were quantified as follows. Geographic coordinates (latitude and longitude) were first determined from existing location data. The environmental variables for each site were subsequently retrieved from the WorldClim database¹²⁴, including altitude (m), temperature ($^{\circ}\text{C}$), solar radiation ($\text{kJ m}^{-2}\text{day}^{-1}$), and water vapor pressure (kPa). Treating environmental factors as quantitative traits, we conducted a GWAS to pinpoint biallelic core SNPs associated with each trait. A total of 250 samples were randomly selected for each trait via a binning-based sampling strategy to ensure uniform coverage of the range of the trait. In brief, trait values were sorted and evenly partitioned into 10 bins, and samples were randomly drawn from each bin. GWAS was conducted via the lin_loc function from the R packages bug (v.0.0.0.9000) and GEMMA (v.0.93)¹²⁵, employing a significance threshold of $-\log_{10}(p\text{-value}) = 5$. The per-site F_{ST} values were computed with the R package PopGenome to reinforce the GWAS findings¹²⁶. The GWAS samples were dichotomized on the basis of trait values for each trait, and this binary classification was used as the basis for the F_{ST} analysis. For the F_{ST} comparisons, the isolates were dichotomized into "high" and "low" groups by selecting the top 125 and bottom 125 samples for each continuous variable.

To identify loci that jointly underlie high-altitude adaptation and clinically relevant phenotypes, we analyzed 239 *H. pylori* isolates with matched clinical metadata from Chengdu and Lhasa (**Extended Data Table 6**) via the same pipeline. Four metrics were considered: altitude-based GWAS scores, altitude F_{ST} , phenotype-based GWAS scores and phenotype F_{ST} . SNPs exceeding the 95th percentile for all four metrics were retained as candidates. Genotypes at each candidate SNP were extracted to construct contingency tables stratified by allele and subgroup (high-altitude cases, high-altitude controls, low-altitude cases and low-altitude controls). Biallelic tables with all expected cell counts ≥ 5 (and with fewer than 20 % of cells < 5) were analyzed by Pearson's χ^2 test; otherwise, Fisher's exact test was applied. ORs comparing high- versus low-altitude allele frequencies were calculated from the corresponding allele contingency tables via Python's statsmodels according to $OR = \frac{(n_A^{case}/n_a^{case})}{(n_A^{control}/n_a^{control})}$, where n_A and n_a denote the counts of alleles A and a, respectively.

To determine whether the effect of genotype on the clinical phenotype differs between high- and low-altitude environments, we fitted, for each SNP, a logistic regression model of the form (phenotype $\sim$ altitude + genotype + altitude $\times$ genotype), extracted the p value for the interaction term, and reported the ten SNPs with the most significant genotype-altitude interactions. Finally, to assess whether a given SNP exerts effects across multiple clinical phenotypes, we first assembled an allele-phenotype matrix for each binary outcome. We then fitted separate logistic regression models of the form $\log \frac{\Pr(Y=1)}{\Pr(Y=0)} = \beta_0 + \beta_{\text{geno}} \times G$, where Y denotes the binary clinical phenotype and where G is the genotype is additively coded. Models were estimated in Python via the statsmodels library,

employing the default Newton algorithm until convergence. ORs were computed as $\exp(\beta_{\text{geno}})$. To correct for multiple testing across m SNPs and k phenotypes, we applied a Bonferroni adjustment, setting the significance threshold at $\alpha = \frac{0.05}{m \times k}$.

Construction of the interactive query system

The system was implemented with the open-source Dash framework, which incorporates the Plotly library to provide dynamic, web-based visualization and user interaction (<https://dash.plotly.com/layout>).

Compliance and ethics

Ethics approval and consent to participate

The Medical Ethics Committee of West China Hospital of Sichuan University approved this study. This study was conducted following the principles of the Helsinki Declaration of 2013.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests. We acknowledge the contributions of Prof. Daniel Falush from the Shanghai Institute of Immunity and Infection at the Chinese Academy of Sciences, the *H. pylori* Genome Project (HpGP) consortium, and others for sharing publicly available *H. pylori* resources.

Funding

This work was supported by the National Natural Science Foundation of China (82402203), the National Natural Science Foundation of Chongqing (CSTB2024NSCQ-LZX0005), the Major Project of the National Social Science Foundation of China (23&ZD203), the Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security (2022FGKFKT05), the Center for Archaeological Science of Sichuan University (23SASA01), the 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYJC20002), and the Sichuan Science and Technology Program (2024NSFSC1518).

Data availability

All data used in this study are provided in the supplementary materials. The all-integrated *H. pylori* genomic resources were submitted to the developed website (<https://www.hpgnomad.top/>). Reference *H. pylori* strains include *H. cetorum* (<https://www.ncbi.nlm.nih.gov/nuccore/CP003479.1/>) and *H. pylori* strain 26695 (https://www.ncbi.nlm.nih.gov/nuccore/NC_000915.1?report=genbank). The public data can be downloaded from databases: NCBI (<https://www.ncbi.nlm.nih.gov/nuccore/?term=h.pylori>), EnteroBase (https://enterobase.warwick.ac.uk/species/helicobacter/search_strains), Datadryad (<https://datadryad.org/stash>) and PubMLST (<https://pubmlst.org/organisms/helicobacter-pylori/>).

Authors' contributions

G.L.H., M.G.W., R.K.T., H.H.D., B.W.Y., H.L., and L.T.L. conceptualized and supervised the project. L.T.L., B.W.L., J.Z., Y.H.L., H.C., L.M.Z., X.Q.T., Y.T.J., Z.H.Z., Y.Z.W., and Y.G.H. conducted data analysis. F.X.B., H.J.Y., C.L., B.W.Y., and H.L. drafted the manuscript, and all the primary authors reviewed, edited, and approved the manuscript.

References

- Valenzuela-Valderrama, M. *et al.* The Helicobacter pylori Urease Virulence Factor Is Required for the Induction of Hypoxia-Induced Factor-1 α in Gastric Cells. *Cancers (Basel)* **11** (2019). <https://doi.org/10.3390/cancers11060799>
- Lamarque, D., Tran Van Nhieu, J. & Breban, M. What are the gastric modifications induced by acute and chronic Helicobacter pylori infection? *Gastroenterol Clin Biol* **27**, 391-400 (2003).
- Mobley, H. L. Helicobacter pylori factors associated with disease development. *Gastroenterology* **113**, S21-28 (1997). [https://doi.org/10.1016/s0016-5085\(97\)80006-6](https://doi.org/10.1016/s0016-5085(97)80006-6)
- Cheok, Y. Y. *et al.* An Overview of Helicobacter pylori Survival Tactics in the Hostile Human Stomach Environment. *Microorganisms* **9** (2021). <https://doi.org/10.3390/microorganisms9122502>
- Kuwahara, H. *et al.* Helicobacter pylori urease suppresses bactericidal activity of peroxynitrite via carbon dioxide production. *Infect Immun* **68**, 4378-4383 (2000). <https://doi.org/10.1128/IAI.68.8.4378-4383.2000>
- Fox, J. G. & Wang, T. C. Inflammation, atrophy, and gastric cancer. *J Clin Invest* **117**, 60-69 (2007). <https://doi.org/10.1172/jci30111>
- Conteduca, V. *et al.* H. pylori infection and gastric cancer: state of the art (review). *Int J Oncol* **42**, 5-18 (2013). <https://doi.org/10.3892/ijo.2012.1701>
- Alvelay, A. O., De Pasos Carrazana, J. L., Cáceres, L. Á., Pagola, M. F. O. & Pérez, A. M. Seguimiento evolutivo mediante estudio endoscópico-histológico e infección por Helicobacter pylori en pacientes con úlcera gástrica. **8**, 20-25 (2010).
- Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **388**, 539-547 (1997). <https://doi.org/10.1038/41483>
- Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513-523 (1991). <https://doi.org/10.1093/genetics/129.2.513>
- Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. Population genetic analysis of Helicobacter pylori by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J Bacteriol* **178**, 3934-3938 (1996). <https://doi.org/10.1128/jb.178.13.3934-3938.1996>
- Moodley, Y. *et al.* Helicobacter pylori's historical journey through Siberia and the Americas. *Proc Natl Acad Sci U S A* **118** (2021). <https://doi.org/10.1073/pnas.2015523118>
- Falush, D. *et al.* Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582-1585 (2003). <https://doi.org/10.1126/science.1080857>
- Linz, B. *et al.* An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915-918 (2007). <https://doi.org/10.1038/nature05562>

Thorell, K. *et al.* The *Helicobacter pylori* Genome Project: insights into *H. pylori* population structure from analysis of a worldwide collection of complete genomes. *Nat Commun* **14**, 8184 (2023). <https://doi.org/10.1038/s41467-023-43562-y>

Tourrette, E. *et al.* An ancient ecospecies of *Helicobacter pylori*. *Nature* **635**, 178-185 (2024). <https://doi.org/10.1038/s41586-024-07991-z>

Moodley, Y. & Linz, B. *Helicobacter pylori* Sequences Reflect Past Human Migrations. *Genome Dyn* **6**, 62-74 (2009). <https://doi.org/10.1159/000235763>

Moodley, Y. *et al.* The peopling of the Pacific from a bacterial perspective. *Science* **323**, 527-530 (2009).

Luo, L. *et al.* Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. *Sci China Life Sci* **68**, 1610-1625 (2025). <https://doi.org/10.1007/s11427-024-2736-7>

Liu, Y., Mao, X., Krause, J. & Fu, Q. Insights into human history from the first decade of ancient human genomics. *Science* **373**, 1479-1484 (2021). <https://doi.org/10.1126/science.abi8202>

Wang, M. *et al.* Multiple Human Population Movements and Cultural Dispersal Events Shaped the Landscape of Chinese Paternal Heritage. *Mol Biol Evol* **41** (2024). <https://doi.org/10.1093/molbev/msae122>

McColl, H. *et al.* The prehistoric peopling of Southeast Asia. *Science* **361**, 88-92 (2018). <https://doi.org/10.1126/science.aat3628>

Yang, M. A. *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282-288 (2020). <https://doi.org/10.1126/science.aba0909>

Yamaoka, Y., Saruuljavkhlan, B., Alfaray, R. I. & Linz, B. Pathogenomics of *Helicobacter pylori*. *Curr Top Microbiol Immunol* **444**, 117-155 (2023). https://doi.org/10.1007/978-3-031-47331-9_5

de Sablet, T. *et al.* Phylogeographic origin of *Helicobacter pylori* is a determinant of gastric cancer risk. *Gut* **60**, 1189-1195 (2011). <https://doi.org/10.1136/gut.2010.234468>

Noto, J. M. *et al.* Pan-genomic analyses identify key *Helicobacter pylori* pathogenic loci modified by carcinogenic host microenvironments. *Gut* **67**, 1793-1804 (2018). <https://doi.org/10.1136/gutjnl-2017-313863>

Yamaoka, Y., Kato, M. & Asaka, M. Geographic differences in gastric cancer incidence can be explained by differences between *Helicobacter pylori* strains. *Intern Med* **47**, 1077-1083 (2008). <https://doi.org/10.2169/internalmedicine.47.0975>

Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021). <https://doi.org/10.3322/caac.21660>

You, Y. *et al.* Genomic differentiation within East Asian *Helicobacter pylori*. *Microb Genom* **8** (2022). <https://doi.org/10.1099/mgen.0.000676>

Suzuki, R. *et al.* *Helicobacter pylori* genomes reveal Paleolithic human migration to the east end of Asia. *iScience* **25**, 104477 (2022). <https://doi.org/10.1016/j.isci.2022.104477>

Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72-75 (2010). <https://doi.org/10.1126/science.1189406>

Wang, M. *et al.* Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm. *The Innovation Life* **3** (2025). <https://doi.org/10.59717/j.xinn-life.2025.100131>

He, G. *et al.* Pilot work of the 10K Chinese People Genomic Diversity Project along the Silk Road suggests a complex east-west admixture landscape and biological adaptations. *Sci China Life Sci* **68**, 914-933 (2025). <https://doi.org/10.1007/s11427-024-2748-4>

Sun, Y. *et al.* Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people. *J Genet Genomics* **51**, 517-530 (2024). <https://doi.org/10.1016/j.jgg.2023.10.002>

Li, X. *et al.* Evolutionary history and biological adaptation of Han Chinese people on the Mongolian Plateau. *hLife* **2**, 296-313 (2024). <https://doi.org/10.1016/j.hlife.2024.04.005>

Wang, M. *et al.* Mighty oaks from little acorns: High-quality genomes of underrepresented populations enhance health equity in precision medicine. *hLife* (2025). <https://doi.org/10.1016/j.hlife.2025.05.014>

He, G. *et al.* Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. *hLife* **2**, 554-562 (2024). <https://doi.org/10.1016/j.hlife.2024.06.006>

Wang, Z. *et al.* YanHuang Paternal Genomic Resource Suggested A Weakly-Differentiated Multi-Source Admixture Model for the Formation of Han's Founding Ancestral Lineages. *Genomics Proteomics Bioinformatics* (2025). <https://doi.org/10.1093/gpbjnl/qzaf049>

Ailloud, F. *et al.* Within-host evolution of *Helicobacter pylori* shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nat Commun* **10**, 2273 (2019). <https://doi.org/10.1038/s41467-019-10050-1>

Berthenet, E. *et al.* A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol* **16**, 84 (2018). <https://doi.org/10.1186/s12915-018-0550-3>

Blanchard, T. G. *et al.* Genome sequences of 65 *Helicobacter pylori* strains isolated from asymptomatic individuals and patients with gastric cancer, peptic ulcer disease, or gastritis. *Pathog Dis* **68**, 39-43 (2013). <https://doi.org/10.1111/2049-632X.12045>

Camorlinga-Ponce, M. *et al.* Phenotypic and Genotypic Antibiotic Resistance Patterns in *Helicobacter pylori* Strains From Ethnically Diverse Population in Mexico. *Front Cell Infect Microbiol* **10**, 539115 (2020). <https://doi.org/10.3389/fcimb.2020.539115>

Chua, E. G. *et al.* Analysis of core protein clusters identifies candidate variable sites conferring metronidazole resistance in *Helicobacter pylori*. *Gastroenterol Rep (Oxf)* **7**, 42-49 (2019). <https://doi.org/10.1093/gastro/goy048>

Didelot, X. *et al.* Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc Natl Acad Sci U S A* **110**, 13880-13885 (2013). <https://doi.org/10.1073/pnas.1304681110>

Gutierrez-Escobar, A. J., Trujillo, E., Acevedo, O. & Bravo, M. M. Phylogenomics of Colombian *Helicobacter pylori* isolates. *Gut Pathog* **9**, 52 (2017). <https://doi.org/10.1186/s13099-017-0201-1>

798 46 Hu, L. *et al.* Long-Read- and Short-Read-Based Whole-Genome Sequencing Reveals the Antibiotic Resistance Pattern of
 799 *Helicobacter pylori*. *Microbiol Spectr* **11**, e0452222 (2023). <https://doi.org/10.1128/spectrum.04522-22>
 800 47 Kojima, K. K. *et al.* Population Evolution of *Helicobacter pylori* through Diversification in DNA Methylation and Interstrain
 801 Sequence Homogenization. *Mol Biol Evol* **33**, 2848-2859 (2016). <https://doi.org/10.1093/molbev/msw162>
 802 48 Mehrotra, T. *et al.* Antimicrobial resistance and virulence in *Helicobacter pylori*: Genomic insights. *Genomics* **113**, 3951-
 803 3966 (2021). <https://doi.org/10.1016/j.ygeno.2021.10.002>
 804 49 Munoz, A. B., Trespalacios-Rangel, A. A. & Vale, F. F. An American lineage of *Helicobacter pylori* prophages found in
 805 Colombia. *Helicobacter* **26**, e12779 (2021). <https://doi.org/10.1111/hel.12779>
 806 50 Munoz-Ramirez, Z. Y. *et al.* Whole Genome Sequence and Phylogenetic Analysis Show *Helicobacter pylori* Strains from
 807 Latin America Have Followed a Unique Evolution Pathway. *Front Cell Infect Microbiol* **7**, 50 (2017).
 808 <https://doi.org/10.3389/fcimb.2017.00050>
 809 51 Munoz-Ramirez, Z. Y. *et al.* A 500-year tale of co-evolution, adaptation, and virulence: *Helicobacter pylori* in the Americas.
 810 *ISME J* **15**, 78-92 (2021). <https://doi.org/10.1038/s41396-020-00758-0>
 811 52 Phuc, B. H. *et al.* *Helicobacter pylori* type 4 secretion systems as gastroduodenal disease markers. *Sci Rep* **11**, 4584 (2021).
 812 <https://doi.org/10.1038/s41598-021-83862-1>
 813 53 Rehvathy, V. *et al.* Multiple Genome Sequences of *Helicobacter pylori* Strains of Diverse Disease and Antibiotic Resistance
 814 Backgrounds from Malaysia. *Genome Announc* **1** (2013). <https://doi.org/10.1128/genomeA.00687-13>
 815 54 Saranathan, R. *et al.* *Helicobacter pylori* Infections in the Bronx, New York: Surveying Antibiotic Susceptibility and Strain
 816 Lineage by Whole-Genome Sequencing. *J Clin Microbiol* **58** (2020). <https://doi.org/10.1128/JCM.01591-19>
 817 55 Shetty, V. *et al.* High primary resistance to metronidazole and levofloxacin, and a moderate resistance to clarithromycin in
 818 *Helicobacter pylori* isolated from Karnataka patients. *Gut Pathog* **11**, 21 (2019). <https://doi.org/10.1186/s13099-019-0305-x>
 819 56 Thorell, K. *et al.* Identification of a Latin American-specific BabA adhesin variant through whole genome sequencing of
 820 *Helicobacter pylori* patient isolates from Nicaragua. *BMC Evol Biol* **16**, 53 (2016). <https://doi.org/10.1186/s12862-016-0619-y>
 821 y
 822 57 Thorell, K. *et al.* Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. *PLoS Genet* **13**, e1006546
 823 (2017). <https://doi.org/10.1371/journal.pgen.1006546>
 824 58 Thorpe, H. A. *et al.* Repeated out-of-Africa expansions of *Helicobacter pylori* driven by replacement of deleterious mutations.
 825 *Nat Commun* **13**, 6842 (2022). <https://doi.org/10.1038/s41467-022-34475-3>
 826 59 Tuan, V. P. *et al.* A Next-Generation Sequencing-Based Approach to Identify Genetic Determinants of Antibiotic Resistance
 827 in Cambodian *Helicobacter pylori* Clinical Isolates. *J Clin Med* **8** (2019). <https://doi.org/10.3390/jcm8060858>
 828 60 Tuan, V. P. *et al.* Genome-wide association study of gastric cancer- and duodenal ulcer-derived *Helicobacter pylori* strains
 829 reveals discriminatory genetic variations and novel oncoprotein candidates. *Microb Genom* **7** (2021).
 830 <https://doi.org/10.1099/mgen.0.000680>
 831 61 Vale, F. F. *et al.* Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins.
 832 *Sci Rep* **7**, 42471 (2017). <https://doi.org/10.1038/srep42471>
 833 62 Zhang, S. *et al.* Mutations of *Helicobacter pylori* RdxA are mainly related to the phylogenetic origin of the strain and not to
 834 metronidazole resistance. *J Antimicrob Chemother* **75**, 3152-3155 (2020). <https://doi.org/10.1093/jac/dkaa302>
 835 63 Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**,
 836 1339+ (2020). <https://doi.org/10.1126/science.aay5012>
 837 64 Arciero, E. *et al.* Whole-genome sequences provide insights into the formation and adaptation of human populations in the
 838 Himalayas. *Curr Biol* **35**, 3624-3637 e3624 (2025). <https://doi.org/10.1016/j.cub.2025.06.048>
 839 65 Bai, F. *et al.* Ancient genomes revealed the complex human interactions of the ancient western Tibetans. *Curr Biol* **34**, 2594-
 840 2605 e2597 (2024). <https://doi.org/10.1016/j.cub.2024.04.068>
 841 66 Wang, M. *et al.* Genomic insights into Neolithic founding paternal lineages around the Qinghai-Xizang Plateau using
 842 integrated YanHuang resource. *iScience* **27**, 111456 (2024). <https://doi.org/10.1016/j.isci.2024.111456>
 843 67 He, G. *et al.* Ancient genomes give insight into 160,000 years of East Asian population dynamics and biological adaptation.
 844 *Genome Biol* (2025).
 845 68 Wood, B. Beachcombing and Coastal Settlement: The Long Migration from South Africa to Patagonia — The Greatest
 846 Journey Ever Made. *Journal of Big History* **3**, 19-46 (2019). <https://doi.org/10.22339/jbh.v3i4.3422>
 847 69 Fernandes, V. *et al.* The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *Am J*
 848 *Hum Genet* **90**, 347-355 (2012). <https://doi.org/10.1016/j.ajhg.2011.12.010>
 849 70 Kong, Q. P. *et al.* Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications
 850 to the peopling of the region. *Mol Biol Evol* **28**, 513-522 (2011). <https://doi.org/10.1093/molbev/msq219>
 851 71 Suzuki, H. & Mori, H. Different Pathophysiology of Gastritis between East and West? An Asian Perspective. *Inflamm Intest*
 852 *Dis* **1**, 123-128 (2016). <https://doi.org/10.1159/000446301>
 853 72 Argent, R. H., Hale, J. L., El-Omar, E. M. & Atherton, J. C. Differences in *Helicobacter pylori* CagA tyrosine phosphorylation
 854 motif patterns between western and East Asian strains, and influences on interleukin-8 secretion. *J Med Microbiol* **57**, 1062-
 855 1067 (2008). <https://doi.org/10.1099/jmm.0.2008/001818-0>
 856 73 Duncan, S. S. *et al.* Comparative genomic analysis of East Asian and non-Asian *Helicobacter pylori* strains identifies rapidly
 857 evolving genes. *PLoS One* **8**, e55120 (2013). <https://doi.org/10.1371/journal.pone.0055120>
 858 74 Yamaoka, Y. *Helicobacter pylori* typing as a tool for tracking human migration. *Clinical Microbiology and Infection* **15**,
 859 829-834 (2009). <https://doi.org/10.1111/j.1469-0691.2009.02967.x>
 860 75 Tomonari, K. *et al.* Fine-scale population structure of Japanese *Helicobacter pylori* provides new anthropological and
 861 epidemiological insights. *Microb Genom* **11** (2025). <https://doi.org/10.1099/mgen.0.001419>
 862 76 Moodley, Y. *et al.* Age of the association between *Helicobacter pylori* and man. *PLoS Pathog* **8**, e1002693 (2012).
 863 <https://doi.org/10.1371/journal.ppat.1002693>

77 Haber, M. *et al.* A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of
 78 Modern Humans Out of Africa. *Genetics* **212**, 1421-1428 (2019). <https://doi.org/10.1534/genetics.119.302368>
 79 Matisoo-Smith, E. & Daugherty, C. Africa to Aotearoa: the longest migration. *Journal of the Royal Society of New Zealand*
 80 **42**, 87-92 (2012). <https://doi.org/10.1080/03036758.2012.673495>
 81 Zhang, X. L. *et al.* The earliest human occupation of the high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago.
 82 *Science* **362**, 1049-1051 (2018). <https://doi.org/10.1126/science.aat8824>
 83 Meyer, M. C. *et al.* Permanent human occupation of the central Tibetan Plateau in the early Holocene. *Science* **355**, 64-67
 84 (2017). <https://doi.org/10.1126/science.aag0357>
 85 Li, Y. C., Tian, J. Y. & Kong, Q. P. A dual origin of Tibetans: evidence from mitochondrial genomes. *J Hum Genet* **60**, 403-
 86 404 (2015). <https://doi.org/10.1038/jhg.2015.40>
 87 Zhao, M. *et al.* Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc*
 88 *Natl Acad Sci U S A* **106**, 21230-21235 (2009). <https://doi.org/10.1073/pnas.0907844106>
 89 Qi, X. *et al.* Genetic Evidence of Paleolithic Colonization and Neolithic Expansion of Modern Humans on the Tibetan Plateau.
 90 *Molecular Biology and Evolution* **30**, 1761-1778 (2013). <https://doi.org/10.1093/molbev/mst093>
 91 Ping, W. J., Xue, J. Y. & Fu, Q. M. Ancient DNA elucidates the migration and evolutionary history of northern and southern
 92 populations in East Asia. *Yi Chuan* **47**, 18-33 (2025). <https://doi.org/10.16288/j.ycz.24-224>
 93 Falush, D. The Remarkable Genetics of *Helicobacter pylori*. *mBio* **13**, e0215822 (2022). <https://doi.org/10.1128/mbio.02158-22>
 94 Shafiee, A., Amini, M., Emamirad, H. & Abadi, A. T. B. Recombination and phenotype evolution dynamics of *Helicobacter*
 95 *pylori* in colonized hosts. *Int J Syst Evol Microbiol* **66**, 2471-2477 (2016). <https://doi.org/10.1099/ijsem.0.001072>
 96 Grant, W. S. Problems and Cautions With Sequence Mismatch Analysis and Bayesian Skyline Plots to Infer Historical
 97 Demography. *J Hered* **106**, 333-346 (2015). <https://doi.org/10.1093/jhered/esv020>
 98 Morelli, G. *et al.* Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS*
 99 *Genet* **6**, e1001036 (2010). <https://doi.org/10.1371/journal.pgen.1001036>
 100 Molak, M., Lorenzen, E. D., Shapiro, B. & Ho, S. Y. Phylogenetic estimation of timescales using ancient DNA: the effects
 101 of temporal sampling scheme and uncertainty in sample ages. *Mol Biol Evol* **30**, 253-262 (2013).
 102 <https://doi.org/10.1093/molbev/mss232>
 103 Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* **23**, 553-559 (2013).
 104 <https://doi.org/10.1016/j.cub.2013.02.044>
 105 Rieun, A. *et al.* Improved calibration of the human mitochondrial clock using ancient genomes. *Mol Biol Evol* **31**, 2780-2792
 106 (2014). <https://doi.org/10.1093/molbev/msu222>
 107 Ermini, L. Out Of Alpine Valleys Through The Time: The Tyrolean Iceman's Complete Mitochondrial Genome And The
 108 Molecular Clock. (2008).
 109 Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of *Mycobacterium tuberculosis*. *PLoS Pathogens*
 110 (2019). <https://doi.org/10.1371/journal.ppat.1008067>
 111 Bai, D. *et al.* Prevalence Difference of *Helicobacter pylori* Infection Between Tibetan and Han Ethnicities in China: A Meta-
 112 analysis on Epidemiologic Studies (SIGES). *Asia Pac J Public Health* **35**, 103-111 (2023).
 113 <https://doi.org/10.1177/10105395221134651>
 114 Bie, M. *et al.* Phylogenetic analysis of clinical strains of *Helicobacter pylori* isolated from patients with gastric diseases in
 115 Tibet. *Ann Transl Med* **7**, 320 (2019). <https://doi.org/10.21037/atm.2019.05.80>
 116 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
 117 data. *Genome Res* **20**, 1297-1303 (2010). <https://doi.org/10.1101/gr.107524.110>
 118 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760
 119 (2009).
 120 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
 121 <https://doi.org/10.1093/bioinformatics/btu153>
 122 Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018).
 123 <https://doi.org/10.1371/journal.pcbi.1005944>
 124 Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037-2039 (2017).
 125 <https://doi.org/10.1093/bioinformatics/btx100>
 126 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
 127 <https://doi.org/10.1093/gigascience/giab008>
 128 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056
 129 (2016). <https://doi.org/10.1099/mgen.0.000056>
 130 Tange, O. *GNU Parallel*, <<https://doi.org/10.5281/zenodo.14207479>> (2024).
 131 Engler, J. B. Tidyploids empowers life scientists with easy code-based data visualization. *Imeta* **4**, e70018 (2025).
 132 <https://doi.org/10.1002/imt2.70018>
 133 Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *The*
 134 *American journal of human genetics* **103**, 338-348 (2018).
 135 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS*
 136 *Genetics* **8**, e1002453 (2012). <https://doi.org/10.1371/journal.pgen.1002453>
 137 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
 138 <https://doi.org/10.1186/s13742-015-0047-8>
 139 Van Rossum, G. a. D., Fred L. *Python 3 Reference Manual*. (CreateSpace, 2009).
 140 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90-95 (2007).
 141 <https://doi.org/10.1109/MCSE.2007.55>

Waskom, M. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021). <https://doi.org/10.21105/joss.03021>

McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018). <https://doi.org/10.21105/joss.00861>

Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026-1028 (2007). <https://doi.org/10.1093/bioinformatics/btm039>

Piñero, C., Abuíñ, J. M. & Pichel, J. C. Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. *Bioinformatics* **36**, 4658-4659 (2020). <https://doi.org/10.1093/bioinformatics/btaa582>

Yu, G. *et al.* ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28-36 (2016). <https://doi.org/10.1111/2041-210x.12628>

Xu, S. *et al.* ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Mol Biol Evol* **38**, 4039-4042 (2021). <https://doi.org/10.1093/molbev/msab166>

Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534 (2020). <https://doi.org/10.1093/molbev/msaa015>

Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**, e1004041 (2015). <https://doi.org/10.1371/journal.pcbi.1004041>

Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007). <https://doi.org/10.1093/molbev/msm088>

Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019). <https://doi.org/10.1371/journal.pcbi.1006650>

Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655-1664 (2009). <https://doi.org/10.1101/gr.094052.109>

Schwarz, S. *et al.* Horizontal versus familial transmission of *Helicobacter pylori*. *PLoS Pathog* **4**, e1000180 (2008). <https://doi.org/10.1371/journal.ppat.1000180>

Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426 (1986). <https://doi.org/10.1093/oxfordjournals.molbev.a040410>

Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000). <https://doi.org/10.1093/oxfordjournals.molbev.a026236>

Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1 - km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302-4315 (2017). <https://doi.org/10.1002/joc.5086>

Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041 (2016). <https://doi.org/10.1038/nmicrobiol.2016.41>

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* **31**, 1929-1936 (2014). <https://doi.org/10.1093/molbev/msu136>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ExtendedDataTable.xlsx](#)
- [ExtendedDataFigs.pdf](#)