Meat without vegetables: An ancient ecotype of *Helicobacter pylori* found in Indigenous populations and animal adapted lineages.

Elise Tourrette[1], Roberto C. Torres[1], Sarah L. Svensson[1], Takashi Matsumoto[2], Muhammad Miftahussurur[3], Kartika Afrida Fauzia[2,3], Ricky Indra Alfaray[2,3], Ratha-Korn Vilaichone[4], Vo Phuoc Tuan[2,5], HelicobacterGenomicsConsortium[#], Abbas Yadegar[6], Lisa Olsson[7], Zhemin Zhou[8], Yoshio Yamaoka[2, 9, 10]*, Kaisa Thorell[11]*, Daniel Falush[1]*

1. The Centre for Microbes, Disease and Health, The research establishment formerly known as Institute Pasteur Shanghai, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai elise.tourrette@gmail.com, rtorres@ips.ac.cn, sarah.svensson@ips.ac.cn, daniel.falush@ips.ac.cn
2. Department of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Yufu, Japan (tmatsumoto9@oita-u.ac.jp) yyamaoka@oita-u.ac.jp
3. Universitas Airlangga, Surabaya, Indonesia (miphto@yahoo.co.id) kartikafauzia@gmail.com rickyindraalfaray@gmail.com
4. Gastroenterology Unit, Department of Medicine, Thammasat University Hospital, Pathum Thani, Thailand. (vilaichone@hotmail.co.th)
5. Cho Ray Hospital, Ho Chi Minh, Vietnam (vophuoctuandr@gmail.com)
6. Foodborne and Waterborne Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran (babak_y1983@yahoo.com)
7. lisa.Olsson@wlab.gu.se University of Gothenburg
8. Pasteurien College, Suzhou Medical College, Soochow University, Suzhou 215123, China zmzhou@suda.edu.cn
9. Department of Medicine, Gastroenterology and Hepatology Section, Baylor College of Medicine, Houston, TX, USA
10. Research center for global and local infectious diseases, Oita University, Yufu, Japan
11. Department of Chemistry and Molecular Biology, Faculty of Science, University of Gothenburg kaisa.thorell@gu.se

[#]See end of manuscript for consortium members.

*Corresponding authors yyamaoka@oita-u.ac.jp , Kaisa.thorell@gu.se, daniel.falush@ips.ac.cn

Abstract
The colonization of our stomachs by *Helicobacter pylori* is believed to predate the oldest splits between extant human populations. We identify a "Hardy" ecotype of *H. pylori* isolated from people belonging to Indigenous populations in Siberia, Canada and USA. The ecotype shares the ancestry of "Ubiquitous" *H. pylori* from the same geographical region in most of the genome but has nearly fixed SNP differences in 100 genes, many of which encode outer membrane proteins and previously identified virulence factors. *H. acinonychis*, found in big cats, and a newly identified primate-associated lineage both belong to the Hardy ecotype and both represent human to animal host jumps. Most strains from the ecotype encode an additional iron-dependent urease that is shared by *Helicobacter* from carnivorous hosts, as well as a tandem duplication of *vacA*, encoding the vacuolating toxin. We conclude that *H. pylori* split into two highly distinct ecotypes approximately 300,000 years ago and that both dispersed around the globe with humans. Coexistence has likely been possible because the ecotypes are adapted to different host diets, but the Hardy ecotype has gone extinct in most parts of the world. Our analysis also pushes back the likely length of the association between *H. pylori* and humans.

Main text
Helicobacter pylori disturbs the stomach lining during long-term colonization of its host [1 2]. Many strains harbour the *cag* pathogenicity island, which encodes a Type IV secretion system that injects a toxin, CagA, into host cells, interfering with cellular signalling and changing their morphology [3]. Individuals infected by *cag*+ strains have been found to have a higher disease burden of gastric cancer than individuals who are uninfected or infected by cag- strains [4]. A second key virulence factor, conserved in all *H. pylori* is VacA, a vacuolating cytotoxin that shows substantial regional variation [1]. However, the range of strategies that *H. pylori* uses and their effect on disease outcomes as well as the important role of host diet remains to be fully characterized [2].

We assembled a dataset of 5626 *H. pylori* and two *H. acinonychis* genome sequences from humans and other hosts around the world. The dataset contains 2663 unpublished genomes from 54 countries, including a large new set of samples from Bhutan, Vietnam, and Iran, as well as strains previously only defined by Multilocus Sequence Typing. Chromosome painting was used to assign the strains to 13 populations and subpopulations, each of which have different geographic distributions (Table S1, S2).

We noticed a surprising discrepancy in clustering between analysis methodologies. A subset of 42 strains from Siberia, Canada and the US formed a clade in phylogenetic trees (Figure S1) and the same strains were separated from all others on a Principal Components Analysis (PCA) plot (Figure S2) and based on pangenome composition. However, the strains were assigned to two different populations, hspSiberia and hspIndigenousNAmerica, with fineSTRUCTURE[5] grouping them with strains from their own geographic regions (Figure S3). These strains were isolated from individuals living in locations that most of the world's population would consider physically inhospitable so we hitherto refer to this clade as "Hardy", while we refer to other *H. pylori* as "Ubiquitous".

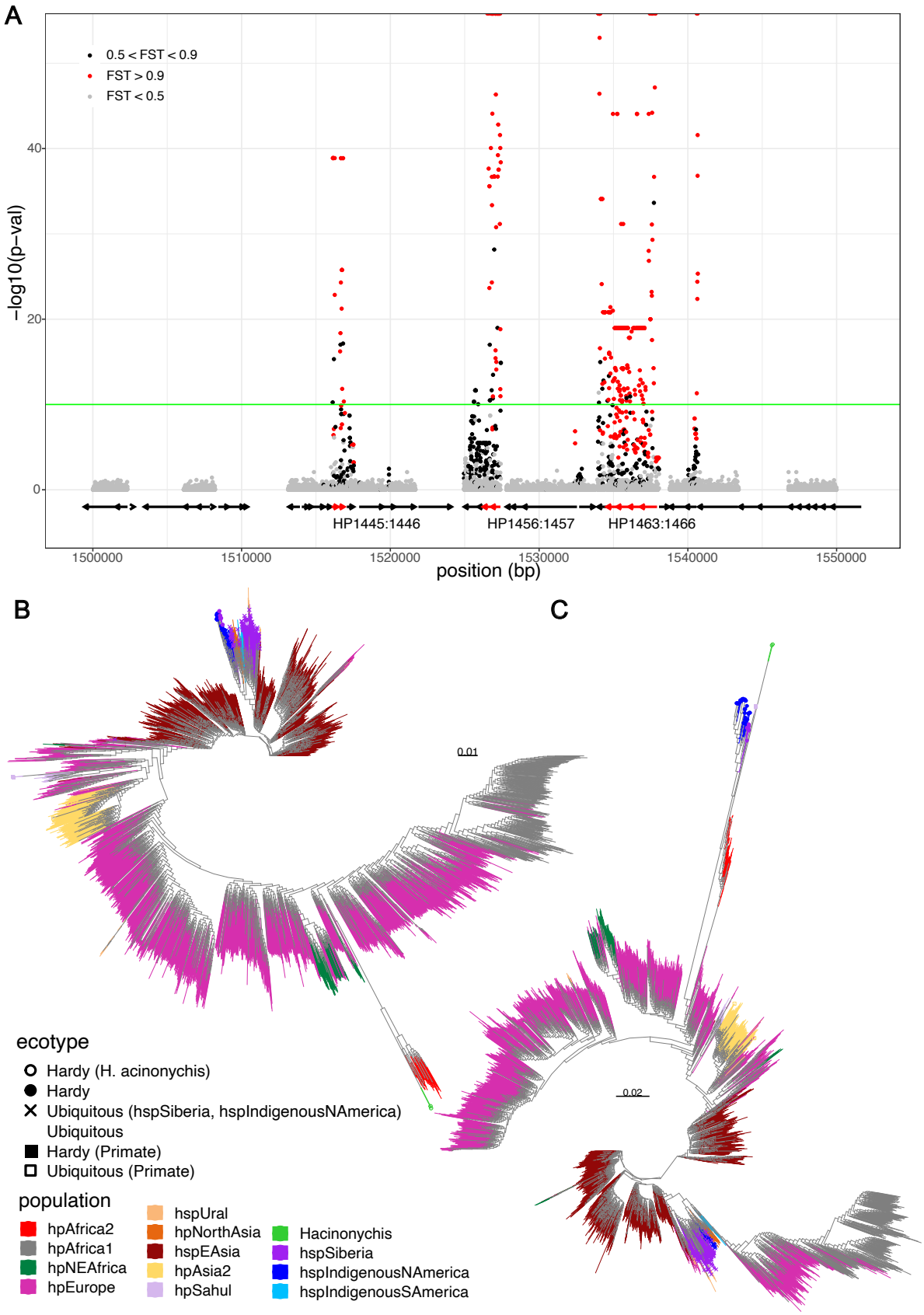

Figure 1:*(A) Manhattan plot resulting from GWAS analysis of the Hardy vs. Ubiquitous strains from hspSiberia and hspIndigenousNAmerica (zoom in between 1.5 and 1.55 Mbp). The genes are indicated by black and red (differentiated) arrows. The green line represents the significance threshold (-log10(p) = 10). The points were coloured based on the F_{ST} values. Phylogenetic tree for the undifferentiated genes (B) and for the differentiated*

genes (C). The branch are colored based on the population and the strains from the Hardy clade are indicated with a dot.

To understand why strains from different locations and ancestry profiles were grouped together by phylogenetic analysis and PCA, we performed a genome-wide analysis (GWAS, genome-wide association study) of differentiation between Hardy clade strains and other strains from the same fineSTRUCTURE populations. The differentiation was localized to specific genomic regions, and was often confined by boundaries of open reading frames (Figure 1A, Figure S4). We therefore split the genome into genes with nearly fixed differences between the Hardy clade and the Ubiquitous clade (100/1577), genes with no differentiated SNPs (1241) and intermediate genes that did not fall into the other two categories (236 genes), to characterize patterns of genetic differentiation separately for these fractions of the genome (Table S3 lists the genes in each category).

Hardy strains have entirely different genetic relationships with other *H. pylori* at differentiated genes compared to undifferentiated ones, as shown by a comparison of phylogenetic trees (Figure 1B, C). The tree of undifferentiated genes (Figure 1B) is consistent with the evolutionary relationships within *H. pylori* established in previous analyses [6 7 8]. Specifically, strains from the hpAfrica2 population are the most genetically distinct human strains. hpAfrica2 strains are associated with Khoisan populations in Southern Africa [7], who also represent the deepest branching population on the human population tree, reflecting a split that took place around 200,000 years ago [9], albeit with considerable uncertainty [10]. hpAfrica2 strains are most closely related to *H. acinonychis*, which has been isolated from tigers and cheetahs in zoos and represents a human-to-animal host jump that has been estimated to have occurred around 50,000 years ago[11]. Other strains cluster into populations whose relationships that reflect human population movements within the last 100,000 years. The Hardy strains cluster with Ubiquitous strains from hspSiberia and hspIndigenousNAmerica populations.

At differentiated genes (Figure 1C), the Hardy strains are the most genetically distinct human strains and cluster with *H. acinonychis* and two *H. pylori* strains that were isolated from primates at the Davis Primate Centre, which has housed rhesus macaques and Cynomolgus monkeys. To reconcile the distinct evolutionary history of these two sets of genes, we hypothesise that *H. pylori* split into two ecotypes prior to the split of hpAfrica2 from other populations (Figure 2), before spreading around the world with humans. The two ecotypes have adapted to different niches, making intermediate genotypes sub-optimal and therefore selecting against recombination between ecotypes at the genes responsible for niche adaptation, but not in the rest of the genome.

The hypothesis implies that the two ecotypes had diverged from each other and were present on both sides of the deepest population split in the history of extant *H. pylori* and humans, which happened in Africa. *H. acinonychis* results from a host jump of an African Hardy strain to big cats, which based on the localized distribution of hpAfrica2 strains almost certainly took place in Africa, while the human Hardy isolates are found in North America and Siberia, and are associated with Indigenous populations (Table S2).

The Davis Primate Center isolates are assigned to the hpSahul population, found in humans in Indonesia, Papua New Guinea, and Australia. They cluster with the hpSahul population in a phylogenetic tree of undifferentiated genes (Figure 1B) as is also confirmed by analysis of genetic distances (Figure S5). Ancestry analysis using ChromoPainter[5] implies that these primate strains fall outside the diversity of currently sampled human hpSahul isolates but are derived from the same ancestral population (Figure S6). Thus, these strains are clearly of human origin, but the timing of the host jump could have been any time within the last 50,000 years. The two strains are highly similar to each other (dS = 0.0192). Two further strains from the UC Davis Primate Research Center belong to the hpAsia2 population and appear to be typical Ubiquitous *H. pylori*. Neither hpSahul nor hpAsia2 are common lineages in the USA (not shown), but are common within the natural host ranges of rhesus macaques and Cynomolgus monkeys in Asia and Oceania suggesting that the host jumps preceded the importation of the primates.

Based on the synonymous divergence of the Hardy clade with other *H. pylori* (dS = 0.342), the Hardy ecotype diverged from other *H. pylori* before the ancestor of hpAfrica2 and *H. acinonychis* (dS = 0.229). Assuming parsimoniously that this divergence happened in humans, this pushes back the association between *H. pylori* and humans to substantially before the first split between extant human populations. This would imply that divergence at the differentiated genes began to accumulate approximately 0.342/0.229*200,000=300,000 years or more ago.

Evidence for atypical patterns of protein evolution at differentiated genes during ecotype divergence comes from patterns of non-synonomous divergence. For Ubiquitous *H. pylori* strains, dS values were similar between differentiated and undifferentiated genes, indicating concordant divergence time for the genes (Figure S7A and B). However, dN/dS values were lower for the differentiated genes, suggesting that these genes are on average under stronger purifying selection than the rest of the genome, consistent with the genes being of high functional importance. The evidence for strong purifying selection contrasts with the elevated rate of protein evolution of Hardy clade strains in comparison with hpAfrica2 strains at the same genes, implying either a loss of constraint or a burst of adaptive evolution in Hardy strains.

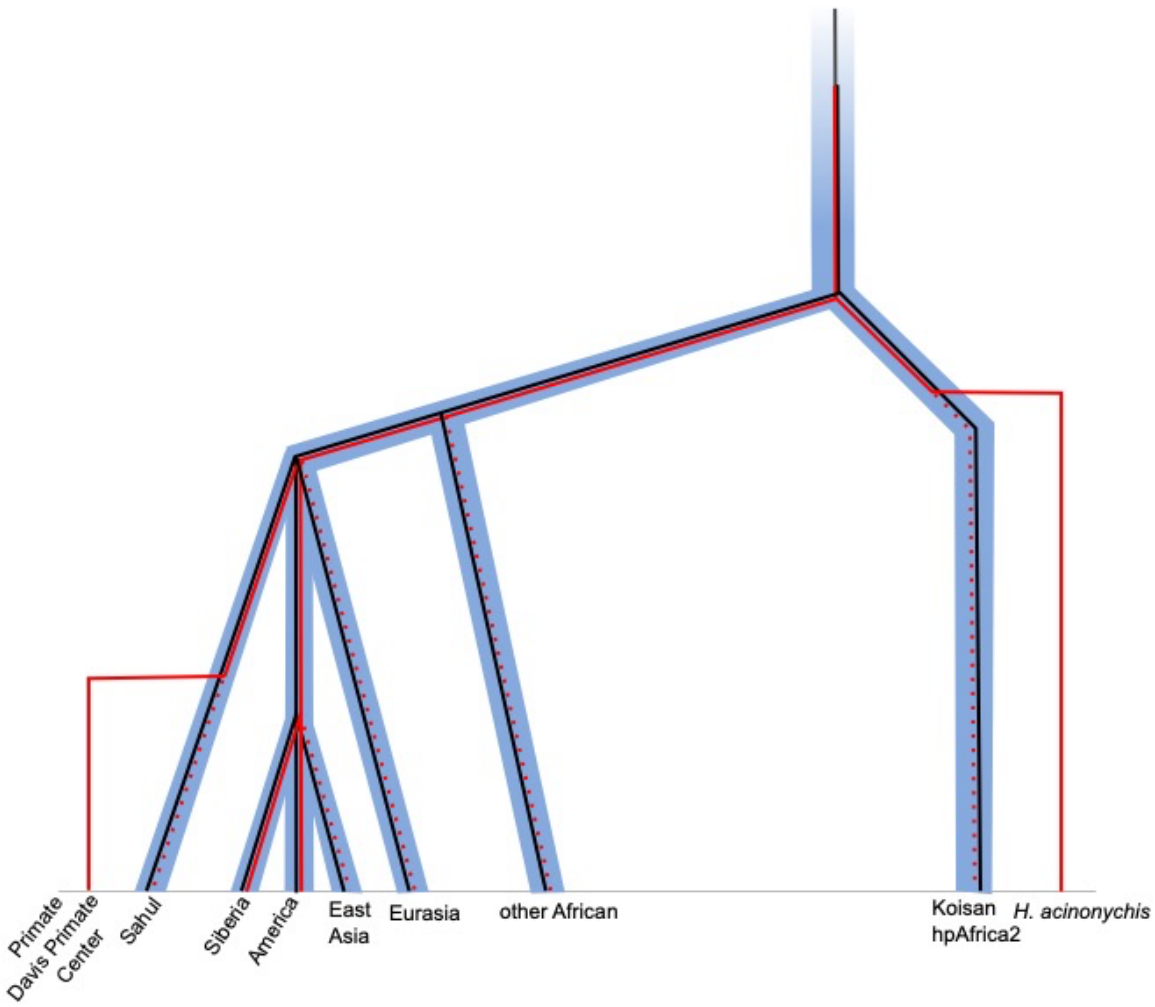

Figure 2*: Scenario for the differentiation of H. pylori into two ecotypes and subsequent global spread. The thick blue tree shows a simplified history of human population differentiation. Helicobacter is represented by thinner lines, which are within the blue tree during periods of evolution that take place within humans. Grey line: H. pylori prior to the evolution of the two ecotypes. Black line: Ubiquitous ecotype. Red line: Hardy ecotype. Dotted red lines indicate that Hardy ecotypes have not yet been detected on the branch and therefore may have gone extinct.*

If natural selection has maintained the differentiation between the two ecotypes, there should be frequent recombination in the bulk of the genome. In line with this prediction, and as described above, strains from the Hardy ecotype have similar ancestry profiles to other strains in the same locations as determined by chromosome painting, which reflects ancestry patterns across the whole genome (Figure S3). Secondly, we should observe gene flow in differentiated genes, but with natural selection acting to keep imported haplotypes at low frequency. The distribution of haplotypes in the highly differentiated regions fits with this prediction (Figure 3A), with sharing of segments between the Hardy ecotype and other strains but with low frequencies for each imported fragment.

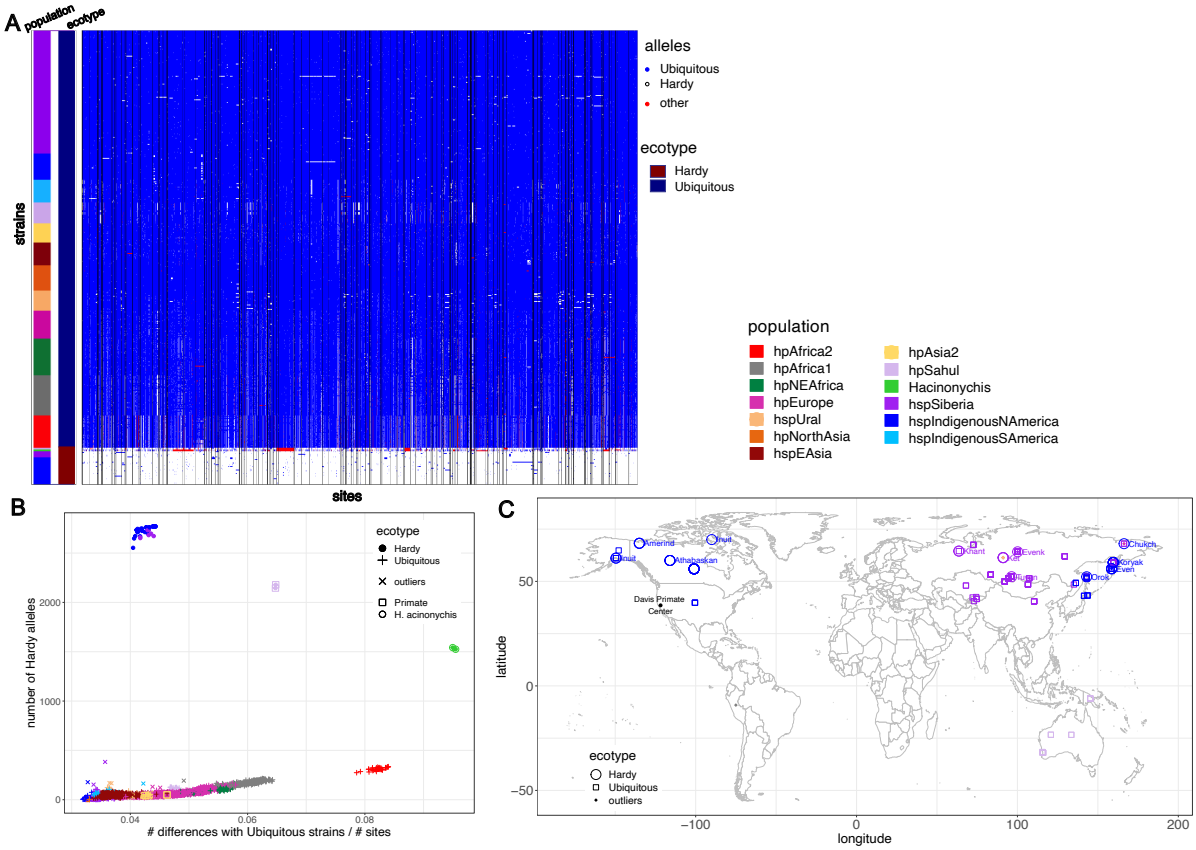

***Figure 3**: (A) Haplotypes at the sites that are differentiated between the two ecotypes for hspSiberia and hspIndigenousNAmerica strains and randomly selected representatives from other populations. The major Hardy allele is represented in white while the major Ubiquitous allele is in blue. Other alleles are shown in red. The black vertical lines separate the different genes. (B) Number of Hardy alleles as a function of average genetic distance with the Ubiquitous ecotype strains (from hspSiberia and hspIndigenousNAmerica). The dots are colored based on their populations and the Hardy strains are represented with a circle. Crosses indicate outlier strains with a higher number of Hardy alleles than expected for a strain at that genetic distance. (C) Map with the location of the Hardy (circle) and Ubiquitous (square) strains from hspSiberia, hspIndigenousNAmerica and hpSahul populations as well as the strains that were outliers (full diamond) compared to their population in the number of Hardy alleles they have.*

Further evidence for selection maintaining the distinction between ecotypes comes from the limited geographic range of haplotypes originating in the Hardy clade. To identify strains that have received such DNA, we plotted the number of Hardy alleles as a function of genetic distance to representatives of the Ubiquitous ecotype from populations where the Hardy ecotype strains were isolated. Distantly related strains such as those belonging to hpAfrica2 have a higher number of such alleles, but this can be attributed to differentiation of the representative Ubiquitous strains rather than to gene flow between ecotypes, as can be seen by the consistent upward trend in Figure 3B. There are nine strains that are clear outliers (Table S4) from this trend line, and all of these come from geographic regions where the Hardy ecotype is found (Figure 3C). The same outliers are identified in a plot of the number of differentiated blocks against the number of differentiated SNPs (Figure S8), showing that the presence of blocks of SNPs characteristic of the other ecotype is also a good marker of recent gene flow. The absence of such blocks in other locations implies that the Hardy ecotype has had a limited recent geographic range. Many of the outlier isolates come from the hspUral population and some from hspIndigenousSAmerica, for which no Hardy genomes have yet been isolated. However, these isolates are not enriched for

Siberian or other ancestry, suggesting that there are likely also Hardy strains within these populations.

The Hardy ecotype involves co-adaptation of more than 100 genes as well as gene presence/absence differences as described below, representing divergent evolutionary strategies. Recombination between ecotypes will create intermediate strains, that are likely to be unfit. Therefore, it has likely become impossible for strains to switch from one strategy to the other, notwithstanding the high level of recent sharing of DNA between the two ecotypes in most of their genome. Consequently, the age of the common clonal ancestor of the two ecotypes should coincide approximately with the origin of the ecotype itself. Concordant with this prediction, as previously noted by other authors for one Hardy strain [12], there are around 140 genome rearrangements differentiating Hardy strains from Ubiquitous strains, contrasting with much lower levels of genome rearrangement within ecotypes. These rearrangements have done little to suppress gene flow between the ecotypes, due to the short length of many fragments incorporated by homologous recombination in *H. pylori* [13].

To investigate the potential role of genetic exchange from other species, we performed BLAST on each of the 100 differentiated genes to identify close matches in a panel of *Helicobacter* genomes from other host species, excluding *H. acinonychis*. In total, 49 genes had significant BLAST matches, the closest of which was always in *H. cetorum*, whose host is dolphins and whales and is the closest relative of *H. pylori* other than *H. acinonychis*. For 28 of the genes, the evolutionary distances, as visualized in a phylogenetic tree (Fig. S9A), were consistent with the scenario shown in Figure 2. For 13 of the genes, the tree instead implied that the Hardy and Ubiquitous versions of the gene diverged from each other after the split between hpAfrica2 and other human *H. pylori*, with hpAfrica2 and *H. acinonychis* versions of the gene clustering together (Fig. S9B). Four genes, *horL, exbB*, a methyl-accepting chemotaxis protein, and *lp220* showed signs of gene flow between Hardy and Ubiquitous versions (Fig. S9C). Three genes, including *vacA* (discussed below) had a history involving gene duplication (Fig. S9D), while for the gene encoding the outer membrane protein HopF, the Hardy version diverged from the Ubiquitous version before splitting from *H. cetorum* (Fig. S9E), which suggests an ancient ancestral polymorphism. There was no gene where the pattern of polymorphism was suggestive of cross-species gene flow.

Clues to the functional basis of ecotype divergence are provided by the genetic composition of strains from the Hardy ecotype. Only five (the two primate strains included) out of 42 have an intact cag pathogenicity island, which is a much lower rate than for the Ubiquitous ecotype (314/478), but enough to imply that the island can function and be maintained by natural selection on both ecotypic backgrounds (Figure 4A and B). The *vacA* gene, coding for vacuolating cytotoxin, is one of the 100 genes differentiated between ecotypes and in addition, one of the Hardy genomes contains an additional tandem pair of duplicated *vacA* genes (Figure 4B and C), while the others contain a large number of fragments, which likely reflect incomplete assembly of repetitive regions within draft genomes. The tandem pair of duplicated *vacA* genes are more distantly related to other human versions of the gene (Fig. S10) and are more closely related to a copy of the gene found in *H. cetorum.* This implies that they diverged from the canonical human copy prior to *H. pylori* splitting into two ecotypes, and suggests that before differentiation into distinct ecotypes occurred, *H. pylori*

genomes may have already contained two distinct copies of the gene, with one being subsequently lost in the Ubiquitous ecotype. However, *H. acinonychis* only carries a single fragmented *vacA* copy [11], while *H. cetorum* from dolphins, on the other hand, carries three divergent *vacA* genes[14]. The phenotypic effect of the second set of *vacA* genes is unclear, although it may serve to increase VacA expression levels or expand the repertoire of VacA alleles.

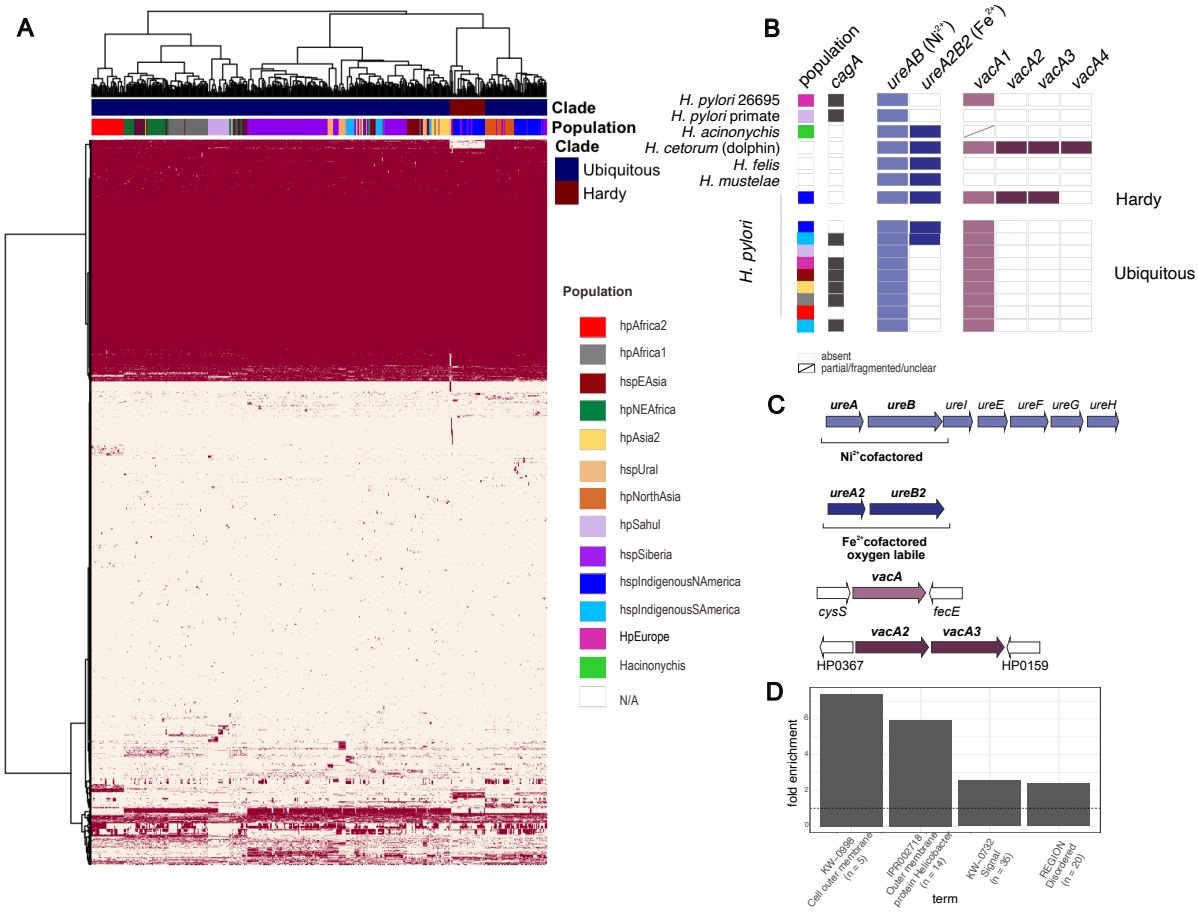

***Figure 4**: (A) Pangenome analysis on a sample of the global dataset: presence in red and absence in white of the different genes. The strains are colored based on their population, their ecotype, and their host species. (B) Presence and absence of cagA and the different types of vacA and ureA/B in a subset of the complete genomes as well as the reference (26695), H. acinonychis, the Hardy primate strains and animal strains were two types of UreA/B have previously been found (H. cetorum, H. felis, H. mustelae). (C) Configuration of the different copies of vacA and ureA/B. (D) Fold enrichment for significant significant (p < 0.05, Fisher's test, Benjamini correction) functional terms.*

Notably, Hardy strains also carry an iron-cofactored urease (UreA2B2) in addition to the canonical nickel-dependent enzyme of *H. pylori* (Fig. 4B and C). Although previously noted in a single Hardy *H. pylori* strain from Northern Canada [12] this feature has otherwise been entirely associated with *Helicobacter* residing in obligate carnivore animal species: *M. mustelae* (ferrets), *H. felis/H. acinonychis* (cats, big cats), and *H. cetorum* (dolphins, whales, pinnipeds) [15]. Iron-cofactored urease is oxygen labile, but thought to allow survival in the low nickel gastric environment of carnivores [16]. Most strains also lack genes encoding AmiE and AmiF, which aid pH buffering in the absence of urea [17]. This points to a markedly different human gastric habitat for the strains. UreA2B2 reflects a very old duplication

event, with versions of both genes also found in *H. cetorum* (Fig. S11 for *ureA* and S12 for *ureB*). Therefore, as for *vacA*, it is likely that the ancestral population had both genes, but that one has been lost by Ubiquitous strains subsequent to the evolution of the ecotype.

Genes in molybdenum-related pathways (Mo uptake, molybdopterin biosynthesis) have decayed in Hardy strains. Loss of these genes was also reported previously for three strains in our dataset [18]. The consequence of this is not clear, as the genes are dispensable for *in vitro* growth of lab *H. pylori*, and the phenomenon is also shared with several East Asian isolates [19].

Functional enrichment analysis of the 100 differentiated genes in the Hardy ecotype confirmed a striking enrichment of *Helicobacter* outer membrane proteins (21/100 genes, $p<0.05$, Fig. 4D) (Table S5 for the list of differentiated genes and their categories, Table S6 for the full result of the functional enrichment analysis). Also enriched were genes encoding proteins 1) with a signal peptide, or 2) located in the inner/outer membrane. These might reflect major structural/functional and/or antigenic differences in the Hardy ecotype cell envelope, as they represent pathways such as iron uptake (TonB1, FrpB4), cell division (TolB/Pal/FtsZ), outer membrane protein secretion/assembly (BamA), outer membrane biogenesis/LPS transport (LptD), and host cell adhesion (AlpB/HopC).

Two transcriptional regulators had fixed genetic differences in the Hardy ecotype: ArsR, which controls a pH-responsive regulon (including the nickel-dependent urease) and Fur, the global iron response regulator. Several additional differentiated genes are also controlled by these regulators (in reference strain 26695). For example, ArsRS regulates several genes that have been lost in the Hardy strains, such as *amiEF*, as well as several fixed genes that it retains (e.g., *hopA*, *horK*, *comH*, and even *fur* itself [20], while Fur controls *vacA* and *tonB* [21] Several are regulated by both ArsR and Fur (e.g., *frpB4, amiE*). *H. pylori* regulators are highly interconnected, and stable differentiation of these transcription factors and their targets between ecotypes suggests these networks have been co-adapted to respond to distinct gastric environments.

Discussion

Our analysis is consistent with two distinct ecotypes of *H. pylori* emerging early in our evolution and spreading around the world in our stomachs. Based on the data presented here, our definition of an ecotype is a distinct strategy encoded by multiple genetic elements that co-occur together on more than one otherwise dissimilar genetic backgrounds. This definition is satisfied because the variants and genes associated with Hardy type are found together in strains from at least three different human *H. pylori* populations as well as *H. acinonychis*. The ecotypes arose long before the split of Khoisan populations from other extant populations, extending the likely length of the association between *H. pylori* and man by at least 50%.

Previously, one set of authors speculated based on a single genome from the Hardy ecotype that its divergent gene content and large number of genome rearrangements might be explained by a bacterium jumping from an animal reservoir to humans [12]. This scenario seems unlikely because it requires at least two host jumps within the last few hundred thousand years, involving unknown reservoir species. Furthermore, the scenario does not provide an obvious explanation for the wide geographical distribution of both ecotypes and the extensive sharing of genetic material between them in most of the genome.

Furthermore, there is no evidence of gene flow from other species occurring at any of the genes that distinguish the two ecotypes. Therefore, it is likely that the divergence and continuing evolution of the ecotype has taken place within humans.

Another set of authors [18] noticed the similarity in gene content between a subset of Hardy hspIndigenousAmerica strains and *H. acinonychis* and suggested that this was due to convergent evolution, involving acquisition of genes from other species. This hypothesis makes intuitive sense because for most of the genome, hspIndigenousAmerica strains cluster with other *H. pylori*, far away from *H. acinonychis*. However, this clustering reflects gene flow. Our analysis implies that *H. acinonychis* shares a common clonal ancestor with other strains from the Hardy ecotype at around the time the Khoisan separated from other human populations. The most recent common ancestor of *any pair* of high and low divergence strains is substantially older than this, as reflected, for example, by the 140 genome rearrangements distinguishing strains even from the same modern *H. pylori* populations[12]. Thus, the gene content shared between Hardy strains, including *H. acinonychis,* arose once and has since been inherited vertically.

Although we do not need to invoke convergent evolution to explain the origin of the two ecotypes, natural selection has played a decisive role in generating and maintaining the distinction between them. There is no evidence of substantial mechanistic barriers to gene flow, despite the many genome rearrangements, meaning that differentiation at each of the 100 genes with nearly fixed differences between ecotypes has been maintained continuously by natural selection in each of the populations where the ecotypes co-occur.

Many bacterial species can colonize a wide range of environments, with adaption often facilitated by large differences in gene complement between strains. However, the extent to which these genes are assembled into stable coadapted gene complexes is not well understood, especially as it can be difficult to disentangle coadaptation from other causes of linkage disequilibrium, such as clonal or population structure. *H. pylori,* as well as free-living *Vibrio parahaemolyticus,* are unusual amongst bacteria for which extensive genomic data is available in that they have undergone sufficient homologous recombination to effectively randomize genetic variation within geographic populations[22]. *V. parahaemolyticus* has a deeply diverged ecotype (EG1a), involving extensive coadaptation between loci encoding lateral flagella and genes involved in cell wall biosynthesis, which is also shared between geographical gene pools and has 18 diverged genome regions. Since both species have clearly defined ecotypes, it is an open question how frequent such ecotypes are within bacteria as a whole and the extent to which their formation and maintenance might be enabled by high rates of homologous recombination.

On the host side, one important implication of our work is that variation in diet has been a continuous feature of human history, with two ecotypes with different dietary capacities both dispersing widely throughout the globe and persisting together stably over long periods in several human populations. However, the Hardy ecotype has gone extinct in most locations. Many details of this long-term coexistence remain to be elucidated, however one parsimonious explanation is that humans relied principally on hunting when colonizing new locations but this depleted large prey [23] and therefore that the first human migrations mainly transmitted the meat-adapted Hardy ecotype, while later migrations brought the

Ubiquitous ecotype, which is better adapted to a broader diet range, including omnivory. The only locations where the Hardy ecotype has been detected so far are habitats where few crops grow or can be easily foraged and people have historically subsisted on a principally hunting-related diet for large parts of the year.

There is an intimate relationship between *H. pylori* pathogenesis and host iron status, and nutritional adaptation has been proposed as one explanation for benefit of virulence factors such as VacA and CagA to the bacterium [2]. The Hardy ecotype has been isolated from human populations with significant gastric disease risk[12] and *H. acinonychis* causes severe gastritis and is a frequent cause of cheetah death in captivity[24]. However, it remains to be seen if the Hardy ecotype can also be defined as a "pathotype". Our hypothesis is that distinct ecotypes have evolved as an adaptation to human dietary variation, however a large fraction of the genes that are differentiated between ecotypes are in outer membrane proteins and other genes associated with virulence. This association is fascinating as it should shed light on the interaction between human diet, bacterial colonization strategies and virulence.

Methods

Genome collection

We collected a total of 8186 *H. pylori* whole genome sequences from public and private sources, including 5508 genomes available in Enterobase[23] (as of Jun 6th 2022); 286 obtained from other public repositories (BIGs and figshare); and 2392 novel strains from different geographic regions around the world, including 266 strains from Iran, collected by Nasrin Mirsaei and Abbas Yadegar and sequenced on the Illumina MiSeq platform at the University of Gothenburg; as well as 2126 genomes from other Asian (1949), African (56), European (5) and American (116) countries, collected by the Department of Environmental and Preventive Medicine, Faculty of Medicine, Oita University and sequenced at Novogene Co., Ltd., Beijing, China with the Illumina NOVA PE150 platform. We filtered out redundant strains from the initial dataset, defined as those sequences with less than 200 SNP distance, and removed low quality genomes based on assembly fragmentation (> 500 contigs), coverage to the 26695 *H. pylori* reference strain (< 70%) and contamination (< 90% *H. pylori*) as predicted by Kraken v2.1.2 [24], to obtain a final total of 5628 non-redundant genomes, correspond to 2663 genomes first reported in this work and 2965 publicly available strains (Table S1)

Genome assembly

Illumina paired-end raw sequence reads obtained from Novogene Co. nwere trimmed with trimmomatic v0.39 [27] and assembled using SPAdes genome assembler v3.15.3 [28] by downsampling read depth to 100 bp, specifying a genome size of 1.6 Mbp and enabling the careful option. Raw reads of the 233 strains sequenced at University of Gothenberg were assembled using the BACTpipe pipeline https://doi.org/10.5281/zenodo.4742358 .

Core Genome variant calling

Single nucleotide polimorphisms from the core genome (core SNPs) were called using a MUMmer v3.20 [29]-based algorithm as previously described [30]. We first aligned each genome sequence of the entire dataset to the *H. pylori* 26695 reference strain (NC_000915.1) using

nucmer v3.1. Then, snp-sites v2.5.2 was used to call all the variants from the whole genome alignment obtained. Variants present in at least 99% of genomes were finally extracted using VCFtools v0.1.17 , obtaining a total of 839,377 core SNPs.

Population assignment

To assign a population to the final dataset, we first defined a reference subset of 285 strains that has been consistently assigned into one of 19 *H. pylori* subpopulations in previous reports [6,8,31,32] and that we confirmed by running fineSTRUCTURE [5] based on haplotype data from core SNPs, computed for this subset as mentioned above, and using 200,000 iterations of burn-in and Markov chain Monte Carlo (MCMC). Then, this subset was considered as a donor panel to paint each sample of the entire dataset using ChromoPainter v2 [5]. Genomes were labeled with a population based on the biggest fraction of their genome painted by a population.

PCA

A PCA (Principal Component Analysis) on the whole dataset was performed using the SNPs extracted from the global alignment file, after linkage disequilibrium-pruning to remove linked SNPs (windows size = 50 bp, step size = 10 variants and r^2 threshold = 0.1), using the software PLINK (v1.9) [33].

Phylogenetic analysis

The different Maximum Likelihood (ML) phylogenetic trees were obtained with FastTree (v2.1.10, option -nt) [34] , based on the nucleotide sequences. When looking at specific genes (*vacA*, *ureA*, *ureB*), the gene sequences were obtained from the individual strains annotation file then aligned using MAFFT (v7.505, option --auto) [35].

FineSTRUCTURE

To further investigate the population structure of Hardy and Ubiquitous strains, we analyzed 209 strains assigned by ChromoPainter to hspSiberia and hspIndigenousNAmerica *H. pylori* populations by running fineSTRUCTURE with 200,000 iterations of burn-in and Markov chain Monte Carlo (MCMC), using as input the haplotype data prepared with SNPs from the core genome of the 209 strains, considering only those variants present in more than 99% of the samples.

GWAS, FST and separation into undifferentiated, intermediate, and differentiated genes

Considering the ecotype as a trait and using the strains from hspSiberia and hspIndigenousNAmerica (209 strains), we performed a GWAS to determine which biallelic core SNPs were significantly associated with the ecotype. HardyHardyThe ubiquitous strains were coded as 0 (171 strains) while Hardy strains were coded as 1 (38 strains). The GWAS was performed using the R [36] package bugwas (v0.0.0.9000) [37] , which considers the population structure via a PCA and then uses GEMMA (v0.93) to perform GWAS analysis. Using a standard significance threshold of -log(p) = 5, we found 3942 significant core biallelic coding SNPs (over 278019 core biallelic SNPs).

To reinforce the results obtained by the GWAS, we calculated the per-site FST between the ubiquitous and Hardy subgroup via the R package PopGenome [38] using the same set of strains and SNPs. We considered a SNP to be differentiated between the Hardy and Ubiquitous subgroups if it was significantly associated with the ecotype by the GWAS (-

log10(p) > 10) and highly differentiated between the two groups based on its FST value (FST > 0.9). We found 2595 differentiated core biallelic coding SNPs (175 differentiated core intergenic SNPs). A SNP was considered to be undifferentiated if its -log(p) was less than 10 and its FST less than 0.5 (257780 CDS SNPs and 8964 intergenic ones). All the other SNPs (7773 CDS SNPs and 607 intergenic) were considered to be intermediate.
Following the separation of the SNPs into three classes, we also distinguished three types of genes: highly differentiated (100 genes, Table S3), with at least 5 differentiated SNPs, lowly/un differentiated (1241 genes), with only undifferentiated SNPs, and the remaining genes (236 genes) being considered as intermediate.

For each strain, we calculated the number of differentiated sites that have the Hardy allele (major allele among the Hardy strains) and compared this number to the genetic distance to the ubiquitous strains and to the number of Hardy block. For a given strain, the distance to the ubiquitous strains is the average number of differences, among SNPs and after removing the gaps, with the ubiquitous strains from hspSiberia and hspIndigenousNAmerica.

Pangenome analysis
First, we estimated the gene content of each sample with prokka v1.4.6 software [39], using the proteome of the 26695 reference strain as reference. Then, .gff files were used as input for Panaroo's v1.2.8[40] pangenome pipeline using the strict mode, merging paralogs based on sequence identity of 0.95, length difference of 0.90 and a core threshold of 0.95. For this analysis smaller dataset was used that consisted of all the strains from hspSiberia and hspIndigenousNAmerica(i.e all the Hardy strains were included and all the ubiquitous strains from the same populations), as well as randomly chosen strains from the other populations (size of the sample dataset: 521 strains)

Comparison to *H. cetorum* and phylogenetic analysis of the differentiated genes
As *H. acinonychis* was considered to be a Hardy strains, we used *H. cetorum* as an outgroup in the study of the differentiated genes. We used the *H. cetorum* strain MIT99-5656 (downloaded the 10/02/2023 from NCBI https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000259275.1/). The gene sequences were obtained from the given Genbank file.
A phylogenetic analysis was performed on the differentiated genes, with the *H. cetorum* genome included. First, we obtained the *H. cetorum* gene sequences by BLASTing (blastn v2.11.0 [41] a Hardy and ubiquitous version of each differentiated gene against the *H. cetorum* genome. For the genes that returned at least one hit, the phylogenetic tree of the gene was generated using FastTree, the *H. cetorum* sequence included.

Functional Enrichment analysis
To determine whether particular types of genes were over-represented among the 100 highly differentiated genes, we performed a functional enrichment analysis, using the website DAVID [42]. Three hypothetical proteins were removed, as they lacked a unique ID. The background set of genes chosen (Table S5 for the list of genes tested and their categories) for the comparison was based on the genes present in *H. pylori* 26695. A term was considered as significant for p > 0.05 after a Benjamini correction for multiple tests.

dN/dS calculations

For each strain in the dataset, we estimated its dN/dS and dS values to hpAfrica2 using the Yang and Nielsen method (YN00) of the PAML (v4.9) software [43 44]. For each strains, their dN/dS and dS value was the average over the values obtained with the different hpAfrica2 strains. The codons that coded for a stop in at least one of the strains were removed from all strains and the dN/dS values were calculated pairwise using the CDS from the global alignment.

References

1. Suerbaum, S. & Michetti, P. Helicobacter pylori infection. *N Engl J Med* **347**, 1175-86 (2002).
2. Amieva, M. & Peek, R.M., Jr. Pathobiology of Helicobacter pylori-Induced Gastric Cancer. *Gastroenterology* **150**, 64-78 (2016).
3. Olbermann, P. *et al.* A global overview of the genetic and functional diversity in the Helicobacter pylori cag pathogenicity island. *PLoS Genet* **6**, e1001069 (2010).
4. Parsonnet, J., Friedman, G.D., Orentreich, N. & Vogelman, H. Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection. *Gut* **40**, 297-301 (1997).
5. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
6. Falush, D. *et al.* Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582-5 (2003).
7. Moodley, Y. *et al.* Age of the association between Helicobacter pylori and man. *PLoS Pathog* **8**, e1002693 (2012).
8. Moodley, Y. *et al.* Helicobacter pylori's historical journey through Siberia and the Americas. *Proc Natl Acad Sci U S A* **118**(2021).
9. Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923-939 e14 (2023).
10. Hollfelder, N., Breton, G., Sjodin, P. & Jakobsson, M. The deep population history in Africa. *Hum Mol Genet* **30**, R2-R10 (2021).
11. Eppinger, M. *et al.* Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet* **2**, e120 (2006).
12. Kersulyte, D. *et al.* Complete Genome Sequences of Two Helicobacter pylori Strains from a Canadian Arctic Aboriginal Community. *Genome Announc* **3**(2015).
13. Kennemann, L. *et al.* Helicobacter pylori genome evolution during human infection. *Proc Natl Acad Sci U S A* **108**, 5033-8 (2011).
14. Kersulyte, D., Rossi, M. & Berg, D.E. Sequence divergence and conservation in genomes of Helicobacter cetorum strains from a dolphin and a whale. *PLoS One* **8**, e83177 (2013).
15. Carter, E.L., Tronrud, D.E., Taber, S.R., Karplus, P.A. & Hausinger, R.P. Iron-containing urease in a pathogenic bacterium. *Proc Natl Acad Sci U S A* **108**, 13095-9 (2011).
16. Stoof, J. *et al.* Inverse nickel-responsive regulation of two urease enzymes in the gastric pathogen Helicobacter mustelae. *Environ Microbiol* **10**, 2586-97 (2008).

17. van Vliet, A.H. *et al.* Differential regulation of amidase- and formamidase-mediated ammonia production by the Helicobacter pylori fur repressor. *J Biol Chem* **278**, 9052-7 (2003).
18. Gressmann, H. *et al.* Gain and loss of multiple genes during the evolution of Helicobacter pylori. *PLoS Genet* **1**, e43 (2005).
19. Kawai, M. *et al.* Evolution in an oncogenic bacterial species with extreme genome plasticity: Helicobacter pylori East Asian genomes. *BMC Microbiol* **11**, 104 (2011).
20. Loh, J.T. *et al.* Delineation of the pH-Responsive Regulon Controlled by the Helicobacter pylori ArsRS Two-Component System. *Infect Immun* **89**(2021).
21. Haley, K.P. & Gaddy, J.A. Metalloregulation of Helicobacter pylori physiology and pathogenesis. *Front Microbiol* **6**, 911 (2015).
22. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. *Bioarxiv* (2019).
23. Dembitzer, J., Barkai, R., Ben-Dor, M. & Meiri, S. Levantine overkill: 1.5 million years of hunting down the body size distribution. *Quaternary Science Reviews* **276**, 107316 (2022).
24. Kusters, J.G., van Vliet, A.H. & Kuipers, E.J. Pathogenesis of Helicobacter pylori infection. *Clin Microbiol Rev* **19**, 449-90 (2006).
25. Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res* **30**, 138-152 (2020).
26. Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).
27. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
28. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-77 (2012).
29. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).
30. Cui, Y. *et al.* Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Mol Biol Evol* **32**, 1396-410 (2015).
31. Linz, B. *et al.* An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915-918 (2007).
32. Moodley, Y. & Linz, B. Helicobacter pylori Sequences Reflect Past Human Migrations. *Genome Dyn* **6**, 62-74 (2009).
33. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
34. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
35. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).
36. Team, R.C. R: A language and environment for statistical computing. (2021).
37. Earle, S.G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041 (2016).
38. 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**, 1929-36 (2014).

39. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9 (2014).
40. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
41. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
42. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
43. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000).
44. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).

HelicobacterGenomicsConsortium

Hafeza Aftab, Department of Gastroenterology, Dhaka Medical College and Hospital , Dhaka, Bangladesh (rosefebruary28@yahoo.com)

Lotay Tshering, Jigme Dorji Wangchuk National Referral Hospital, Thimphu, Bhutan, lotay444@gmail.com

Dhakal Guru Prasad, Jigme Dorji Wangchuk National Referral Hospital, Thimphu, Bhutan, (gpdhakal@jdwnrh.gov.bt)

Evariste Tshibangu-Kabamba, University of Mbujimayi, Mbujimayi, DR Congo / Osaka Metropolitan University, Osaka, Japan (evaristetshibangu@gmail.com)

Ghislain Disashi Tumba, University of Mbujimayi, Mbujimayi, DR Congo (tdisashi@yahoo.fr)

Patrick de Jesus Ngoma-Kisoko, University of Kinshasa, Kinshasa, DR Congo (patrickdejesus3@gmail.com)

Antoine Tshimpi-Wola, University of Kinshasa, Kinshasa, DR Congo (antshimpi@aol.com)

Dieudonné Mumba Ngoyi, University of Kinshasa, Kinshasa, DR Congo / National Institute of Biomedical Research (INRB), DR Congo (mumbadieudonne@yahoo.fr)

Pascal Tshiamala Kashala, Gastroenterology Service, Astryd Clinics, Kinshasa, DR Congo (ptshiamala@yahoo.fr)

Modesto Cruz, Instituto de Microbiología y Parasitología (IMPA), Universidad Autónoma de Santo Domingo (UASD), Santo Domingo (mcruz30@uasd.edu.do)

José Jiménez Abreu, Instituto de Microbiología y Parasitología (IMPA), Universidad Autónoma de Santo Domingo (UASD), Santo Domingo (jojis17@gmail.com)

Celso Hosking, Universidad Autónoma de Santo Domingo, Santo Domingo (celsojnc@gmail.com)

Hideo Terao, Oita University, Oita, Japan (teraohdr@gmail.com)

Titong Sugihartono, Universitas Airlangga, Surabaya, Indonesia (titongsppd@gmail.com)

Ari Fahrial Syam, University of Indonesia, Jakarta, Indonesia (ari_syam@hotmail.com)

Langgeng Agung Waskito, Universitas Airlangga, Surabaya, Indonesia (langgengaw@gmail.com)

Hasan Maulahela, University of Indonesia, Jakarta, Indonesia (hasan.maulahela@yahoo.com)

Yudith Annisa Ayu Rezkitha, Faculty of Medicine, Muhammadiyah University of Surabaya, Surabaya, Indonesia (yudithannisaayu@gmail.com)

Thein Myint, , Department of Gastroenterology, Yangon General Hospital, University of Medicine, Yangon, Myanmar.(theinmyint06@gmail.com)

Than Than Aye, Department of Gastroenterology, Thingangyun Sanpya General Hospital, University of Medicine (2), Thingangyun, Myanmar

New Ni, Department of Gastroenterology, Mandalay General Hospital, University of Medicine, Mandalay, Myanmar.

Than Than Yee, Department of GI and HBP Surgery, No. (1) Defense Service General Hospital (1000 Bedded), Mingaladon, Yangon 11021, Myanmar (drthanthanyee@gmail.com)

Kyaw Htet, Defence Services General Hospital, Yangon, Myanmar (drkyawhtet@gmail.com)

Takeshi Matsuhisa, Nippon Medical School, Tokyo, Japan (matuhisa@m8.dion.ne.jp)

Pradeep Krishna Shrestha, Tribhuvan University Teaching Hospital (drpkshrestha@hotmail.com)

Rabi Prakash Sharma, Gastroenterology Department, Maharajgunj Medical Campus, Tribhuvan University Teaching Hospital, Kathmandu, Nepal.(rabiprakash2001@yahoo.com)

Jeewantha Rathnayake, Department of Surgery, Teaching Hospital Peradeniya, University of Peradeniya, Kandy 20404, Sri Lanka (jeewanrath@gmail.com)

Meegahalande Durage Lamawansa, Department of Surgery, Teaching Hospital Peradeniya, University of Peradeniya, Kandy 20404, Sri Lanka (mdyasas@yahoo.co.uk)

Varocha Mahachai, GI and Liver Center, Bangkok Medical Center, Bangkok, Thailand (varocha.m@gmail.com)

Thawee Ratanachu-Ek, Department of Surgery, Rajavithi Hospital, Bangkok, Thailand (thawee1958@gmail.com)

Phawinee Subsomwong, Department of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Yufu, Japan. Department of Microbiology and Immunology, Hirosaki University Graduate School of Medicine, Hirosaki 036-8562, Japan. phawinee@hirosaki-u.ac.jp

Tran Thi Huyen Trang, 108 Military Central Hospital, Hanoi, Vietnam (huyentrang110@yahoo.com)

Tran Thanh Binh, Cho Ray Hospital, Ho Chi Minh, Vietnam (bs.binh@yahoo.com.vn)

Ho Dang Quy Dung, Cho Ray Hospital, Ho Chi Minh, Vietnam (quydung@gmail.com)

Vu Van Khien, 108 Military Central Hospital, Hanoi, Vietnam
(vuvankhien108@yahoo.com.vn)

Dou Narith, Department of Endoscopy, Cho Ray Phnom Penh Hospital, Phnom Penh 12357, Cambodia. dr.narith@yahoo.com

Shaho Negahdar Panirani, MSc, Foodborne and Waterborne Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. (shaho.n.p1990@gmail.com)

Hamid Asadzadeh Aghdaei, MD, Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. (hamid.assadzadeh@gmail.com)

Mohammad Reza Zali, MD, Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
(nnzali@hotmail.com)

Nasrin Mirzaei, MSc, Foodborne and Waterborne Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
(mirzaei.nasrin@yahoo.com)

Emilio Rudbeck, MSc, Clinical Genomics Gothenburg, Bioinformatics and Data Centre, University of Gothenburg, Sweden (emilio.rudbeck@gu.se)

Acknowledgements
We thank Mark Achtman for providing strain DNA and Jay Solnick for information on primate strains.

Funding:

This work is supported by National Natural Science Foundation of China (No. 32170640, No. 32211550014) to Daniel Falush and by Shanghai Municipal Science and Technology Major Project No. 2019SHZDZX02. The work is also supported by funding from Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan (18KK0266, 19H03473, 21H00346 and 22H02871) (Yoshio Yamaoka) and 21K08010 (Takashi Matsumoto) and by the Grants-in-aid of the National Fund for Innovation and Development of Science and Technology (FONDOCYT) from the Ministry of Higher Education Science and Technology of the Dominican Republic (MESCyT), No. 2012-2013-2A1-65 and No. 2015-3A1-182 (Modesto Cruz). This work was also supported by the Japan Agency for Medical Research and Development (AMED)(e-ASIA JRP) and Japan Society

for the Promotion of Science (JSPS) (Bilateral Programs; Japan-China) (Yoshio Yamaoka), as well as by the Thailand Science Research and Innovation Fundamental Fund, Bualuang ASEAN Chair Professorship at Thammasat University, and Center of Excellence in Digestive Diseases, Thammasat University, Thailand (Ratha-korn Vilaichone). Kaisa Thorell was supported by Swedish Society for Medical Research (SSMF), Assar Gabrielsson Foundation FB20-12 and FB21-89, and Magnus Bergvall Foundation. The computations and data storage required for assembly and annotation of genomes sequenced at University of Gothenburg were enabled by resources in projects snic-2021/22-229 and snic-2021/23-234 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at the UPPMAX HPC, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.