**An ancient ecospecies 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[#], Difei Wang[6], Abbas Yadegar[7], Lisa M. Olsson[8], Zhemin Zhou[9], Yoshio Yamaoka[2, 10, 11]*, Kaisa Thorell[12]*, Daniel Falush[1]*

1. The Centre for Microbes, Disease and Health, 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. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20850, USA. difei.wang2@nih.gov
7. 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
8. The Wallenberg Laboratory, Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Sweden. lisa.olsson@wlab.gu.se
9. Pasteurien College, Suzhou Medical College, Soochow University, Suzhou 215123, China. zmzhou@suda.edu.cn
10. Department of Medicine, Gastroenterology and Hepatology Section, Baylor College of Medicine, Houston, TX, USA.
11. Research center for global and local infectious diseases, Oita University, Yufu, Japan.
12. Department of Chemistry and Molecular Biology, Faculty of Science, University of Gothenburg, Sweden. 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" ecospecies of *H. pylori* associated with indigenous groups, isolated from people in Siberia, Canada, and USA. The ecospecies 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 host interaction factors. For these parts of the genome, the ecospecies has a separate, independently evolving gene pool with a distinct evolutionary history. *H. acinonychis*, found in big cats, and a newly identified primate-associated lineage both belong to the Hardy ecospecies and both represent human to animal host jumps. Most strains from the ecospecies 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 ecospecies approximately 300,000 years ago and that both dispersed around the globe with humans. Coexistence has likely been possible because the ecospecies are adapted to different host diets, but the Hardy ecospecies 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 host interaction factor, conserved in all *H. pylori,* is VacA, a vacuolating cytotoxin that shows substantial geographical variation (*1*).
These strategies, and others that remain to be determined, are likely a response of the bacteria to changes in its environment, such as the diets of their hosts.
We would therefore like to characterize global variation in *H. pylori* genetic strategies and link that variation to environmental factors. We will then be better positioned to understand the nature of interactions between environment, host and bacteria that lead to the wide range of clinical outcomes in infected individuals (*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 Southeast Asia, and Iran, as well as strains previously only defined by Multilocus Sequence Typing, among them a high number of genomes from Siberia. Chromosome painting was used to assign the strains to 13 populations and subpopulations, each of which have different geographic distributions (Table S1).

We noticed that different methods for clustering strains gave surprisingly different answers. 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, and

fineSTRUCTURE (*5*) grouped 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".

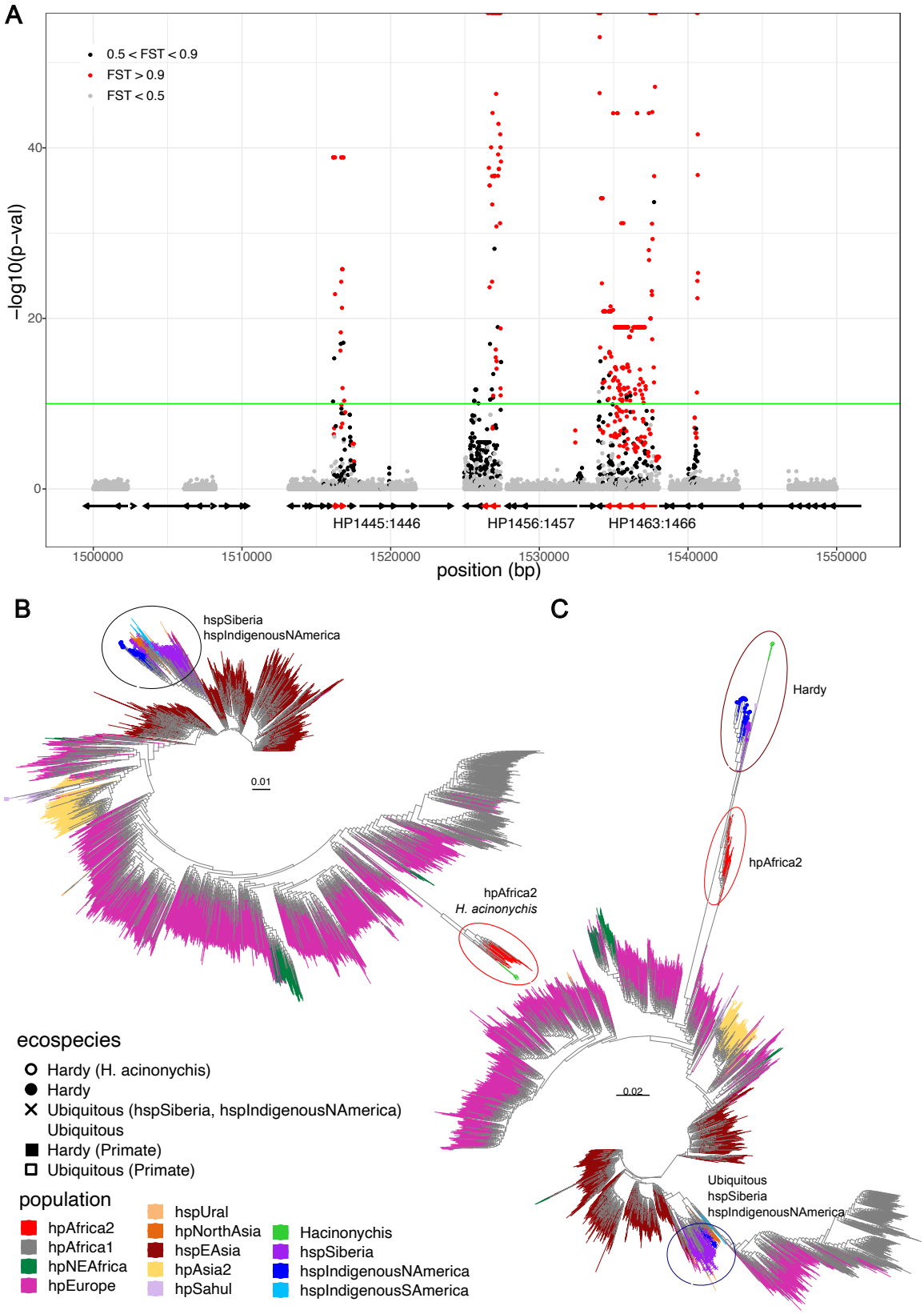

***Figure 1***:*(A) Manhattan plot resulting from GWAS analysis of the Hardy vs. Ubiquitous strains from hspSiberia and hspIndigenousNAmerica (zoom in between 1.5 and 1.55 Mbp). The genes are indicated by black and red (differentiated) arrows. The green line represents the significance threshold (-log10(p) = 10). The points were coloured based on the $F_{ST}$ values. Half dots at the top of the plot indicate an estimated P-value of zero and $F_{ST}$ of one. Phylogenetic trees 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 gene boundaries (Figure 1A, Figure S4). We therefore split the genome alignment into genes with nearly fixed differences between the Hardy clade and the Ubiquitous clade (100/1577), genes with no differentiated SNPs (1048) and intermediate genes that did not fall into the other two categories (429 genes) to characterize patterns of genetic differentiation separately for these fractions of the genome (Table S2 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 (*6-8*), with Hardy strains from hspSiberia and hspIndigenousNAmerica populations clustering with Ubiqutous strains isolated from the same locations. The deepest branch in the tree separates strains from hpAfrica2 population and *H. acinonychis* from others *H. pylori*. hpAfrica2 strains are associated with Khoisan populations in Southern Africa (*7*), who also represent the deepest branch on the human population tree, reflecting a split that took place around 200,000 years ago (*9*), albeit with considerable uncertainty (*10*). *H. acinonychis*, which has been isolated from big cats in zoos, represents a human-to-animal host jump that has been estimated to have occurred at least 50,000 years ago (*11*).

At differentiated genes (Figure 1C), the Hardy strains isolated from humans are the most genetically distinct *H. pylori*, branching more deeply than hpAfrica2. The Hardy strains cluster with *H. acinonychis* and two *H. pylori* strains that were isolated from primates at the Davis Primate Center, which has housed rhesus macaques and Cynomolgus monkeys. These isolates are assigned to the hpSahul population, found in humans in Indonesia, Papua New Guinea, and Australia and cluster with with other isolates from the population in the tree of undifferentiated genes (Figure 1B). However, the ancestry profile of these isolates is clearly distinct from those of any other hpSahul isolate in the database (Figure S5), suggesting the host jump may have been ancient. Two further strains from the UC Davis Primate Research Center belong to the hpAsia2 population and appear to be typical Ubiquitous *H. pylori*.

To reconcile the distinct evolutionary history of these two sets of genes, we hypothesise that *H. pylori* split into two ecospecies prior to the split of hpAfrica2 from other populations (Figure 2A). We define ecospecies as bacteria within the same species that have adapted to different niches and thus have acquired distinct gene pools in a sizeable fraction of the

genome, due to natural selection against intermediate genotypes in that fraction. The two ecospecies were both spread around the world by human migrations. Since the gene pools at differentiated loci are distinct, genetic relationships at these genes represent the separate histories of spread of each ecospecies, which may have been carried to their present locations by different human migrations, while undifferentiated genes reflect an amalgum of both histories. Phylogenetic trees constructed for Hardy and Ubiquitous strains from the same populations at differentiated genes show similar, parallel, histories (Figure 2B).

To estimate the age of the ecospecies, we need a consistent molecular clock. dS levels are similar for ubiquitous strains in differentiated and undifferentiated regions (Figure S6A and B), implying no large difference in mutation rates between the two regions. In differentiated regions, comparisons between ecospecies gave divergence levels greater than 0.3 for all comparisons involving *H. acinonychis* or hpAfrica2 (Figure S6B and D), implying that the ecotypes arose substantially before the divergence of hpAfrica2 from other *H. pylori* (dS = 0.23). Comparisons involving other Hardy strains showed intermediate dS values, with the lowest values for strains where Hardy and Ubiquitous coexist, which we attribute to residual gene flow even within the differentiated regions, which, as expected, is weaker at non-synonymous sites. Based on dS values between hpAfrica2 and *H. pylori* Hardy strains, divergence between ecospecies at the differentiated genes began to accumulate approximately 0.34/0.23*200,000=300,000 years or more ago.

dN/dS levels indicate the degree of functional constraint on genes and are also impacted by positive selection (Figure S6A and B). Within Ubiquitous *H. pylori*, dN/dS levels are lower in differentiated genes than in undifferentiated ones, implying than differentiated genes are on average more functionally constrainted than the rest of the genome. However, dN/dS levels between Hardy and hpAfrica2 strains are elevated in differentiated genes, which implies that after the differentiation between the two ecospecies there was a burst of evolution at functional sites at these genes.

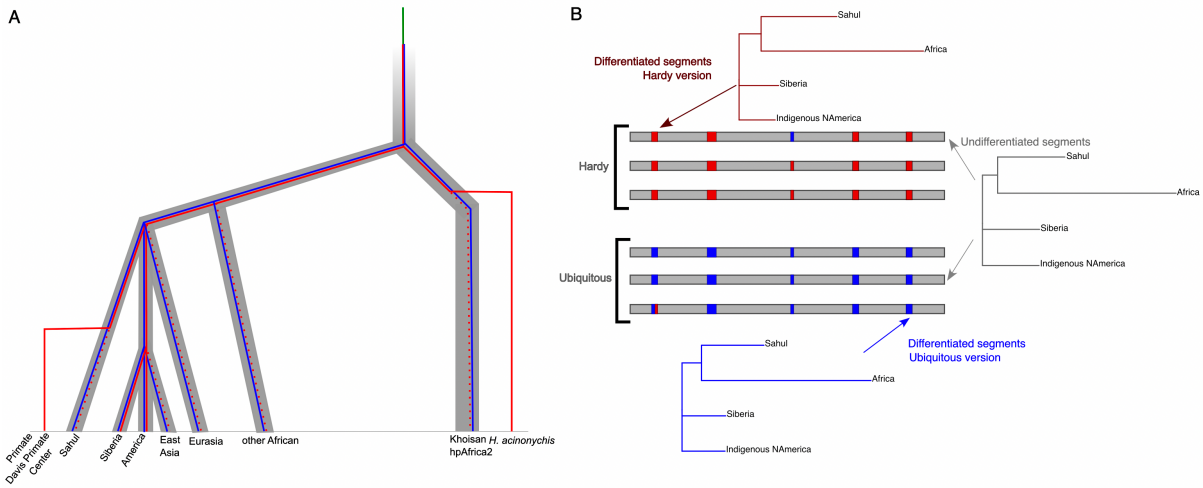


***Figure 2**: (A) Scenario for the differentiation of Helicobacter pylori into two ecospecies and subsequent global spread. The thick grey tree shows a simplified history of human population differentiation. Helicobacter is represented by thinner lines, which are within the grey tree during periods of evolution that take place within humans. Green line: H. pylori prior to the evolution of the two ecospecies. Blue line: Ubiquitous ecospecies. Red*

*line: Hardy ecospecies. Dotted red lines indicate that Hardy ecospecies have not yet been detected on the branch and therefore may have gone extinct. (B) Phylogenetic trees for the differentiated (Hardy, in red, and Ubiquitous, in blue) and undifferentiated (in grey) regions of the genomes based from the genomes from hpAfrica2 and H. acinonychis for Africa, hpSahul, hspSiberia, hspIndigenousNAmerica.*

The fineSTRUCTURE results demonstrate that there is frequent recombination between ecospecies in most of the genome (Figure S3). The question is then what force has maintained the differentiated regions. If this force is natural selection, then we should see genetic exchange between ecospecies that is too recent to have been purged. The distribution of haplotypes in the differentiated regions fits with the action of natural selection (Figure 3A), with sharing of segments between the Hardy ecospecies and other strains in many genes but with low frequencies for each imported fragment.

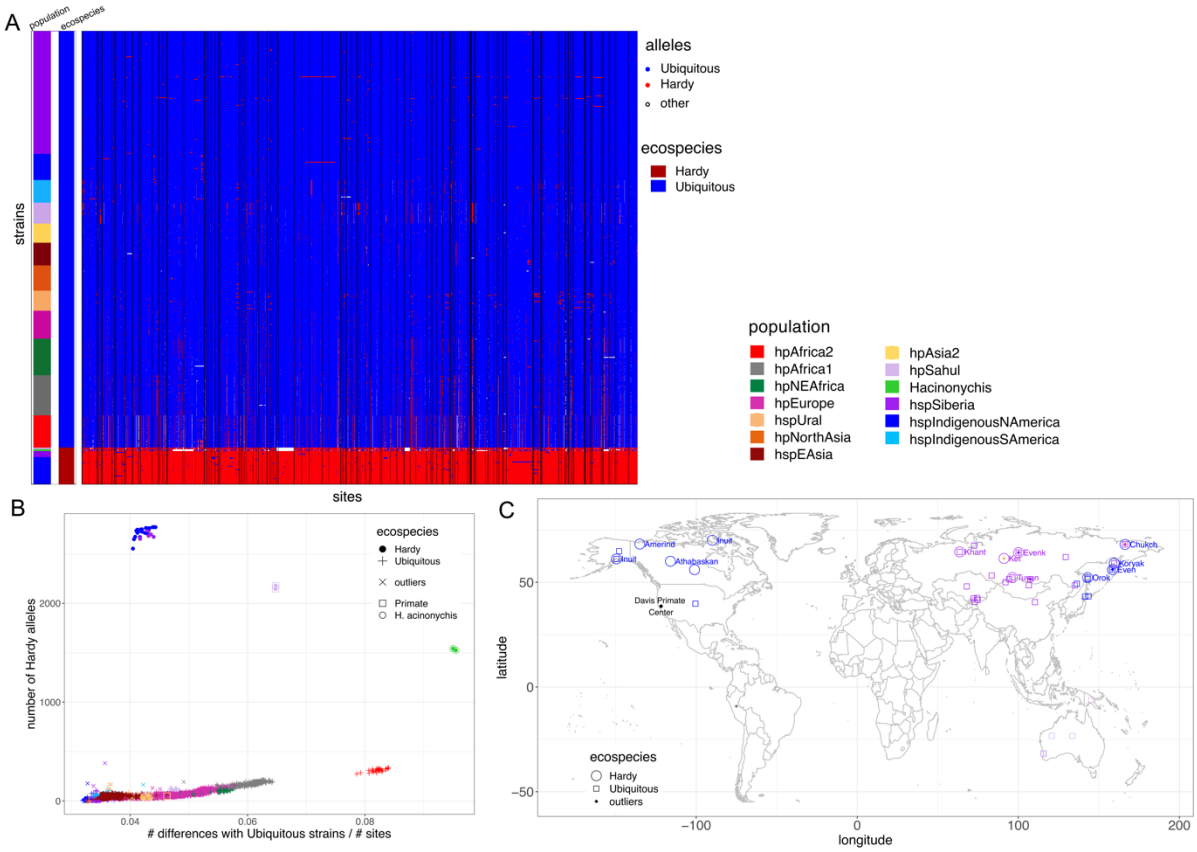


***Figure 3**: (A) Haplotypes at the sites that are differentiated between the two ecospecies for hspSiberia and hspIndigenousNAmerica strains and randomly selected representatives from other populations. The major Hardy allele is represented in red while the major Ubiquitous allele is in blue. Other 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 ecospecies strains (from hspSiberia and hspIndigenousNAmerica). The dots are coloured 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 ecospecies 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, i.e. the predominant allele in the Hardy clade, as a function of genetic distance to representatives of the Ubiquitous ecospecies from populations where the Hardy ecospecies 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 ecospecies, as can be seen by the consistent upward trend in Figure 3B. There are nine strains that are clear outliers (Table S3) from this trend line, and most of these come from geographic regions where the Hardy ecospecies 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 S7), showing that the presence of blocks of SNPs characteristic of the other ecospecies is also a good marker of recent gene flow. The absence of such blocks in other locations implies that the Hardy ecospecies has had a limited recent geographic range. Many of the outlier isolates come from the hspUral population and one 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 ecospecies involves co-adaptation of more than 100 genes as well as gene presence/absence differences as described below, representing divergent evolutionary strategies. Recombination between ecospecies 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 ecospecies in most of their genome. Consequently, the age of the common clonal ancestor of the two ecospecies should coincide approximately with the origin of the ecospecies itself.

Recombination between strains transfers DNA but is less likely to affect genome order, especially for large rearrangements. Therefore, rearrangements can potentially provide a marker of clonal age, even in the presence of high recombination rates. Dotplots revealed few rearrangements between genomes from the same ecospecies, including between *H. acinonychis* and other Hardy strains, albeit more distantly related genomes had more rearrangements (Figure S8). However there was a considerably larger number of rearrangements between strains from different ecospecies, even when the strains came from the same fineSTRUCTURE population. These rearrangements have done little to suppress gene flow between the ecospecies, due to the short length of many fragments incorporated by homologous recombination in *H. pylori* (*12*).

Molecular clocks could be misleading if Hardy variants were obtained by horizontal genetic exchange with other *Helicobacter* species. To investigate this, 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 hosts are 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 (Figure 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 HP0599, and *lp220*) showed signs of gene flow between Hardy and Ubiquitous versions (Figure S9C). Three genes, including *vacA* (discussed below) had a history involving gene duplication (Figure S9D), while for the gene encoding the outer membrane protein HopF, the Hardy version diverged from the Ubiquitous version before splitting from *H. cetorum* (Figure S9E). This pattern could be explained by cross species genetic exchange but an alternative is that the common ancestor of *H. pylori* and *H. cetorum* had two different versions of the gene, with both copies being maintained in *H. pylori* until it split into ecospecies. There was no other gene where the pattern of polymorphism was suggestive of import of genetic material from other *Helicobacter* species.

Clues to the functional basis of ecospecies divergence are provided by the genetic composition of strains from the Hardy ecospecies. 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 ecospecies (314/478 strains), 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 ecospecies and in addition, the completely sequenced Hardy genomes reveal 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 (Figure 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 ecospecies, and suggests that before differentiation into distinct ecospecies occurred, *H. pylori* genomes may have already contained two distinct copies of the gene, with one being subsequently lost in the Ubiquitous ecospecies. 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 (*13*). 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.

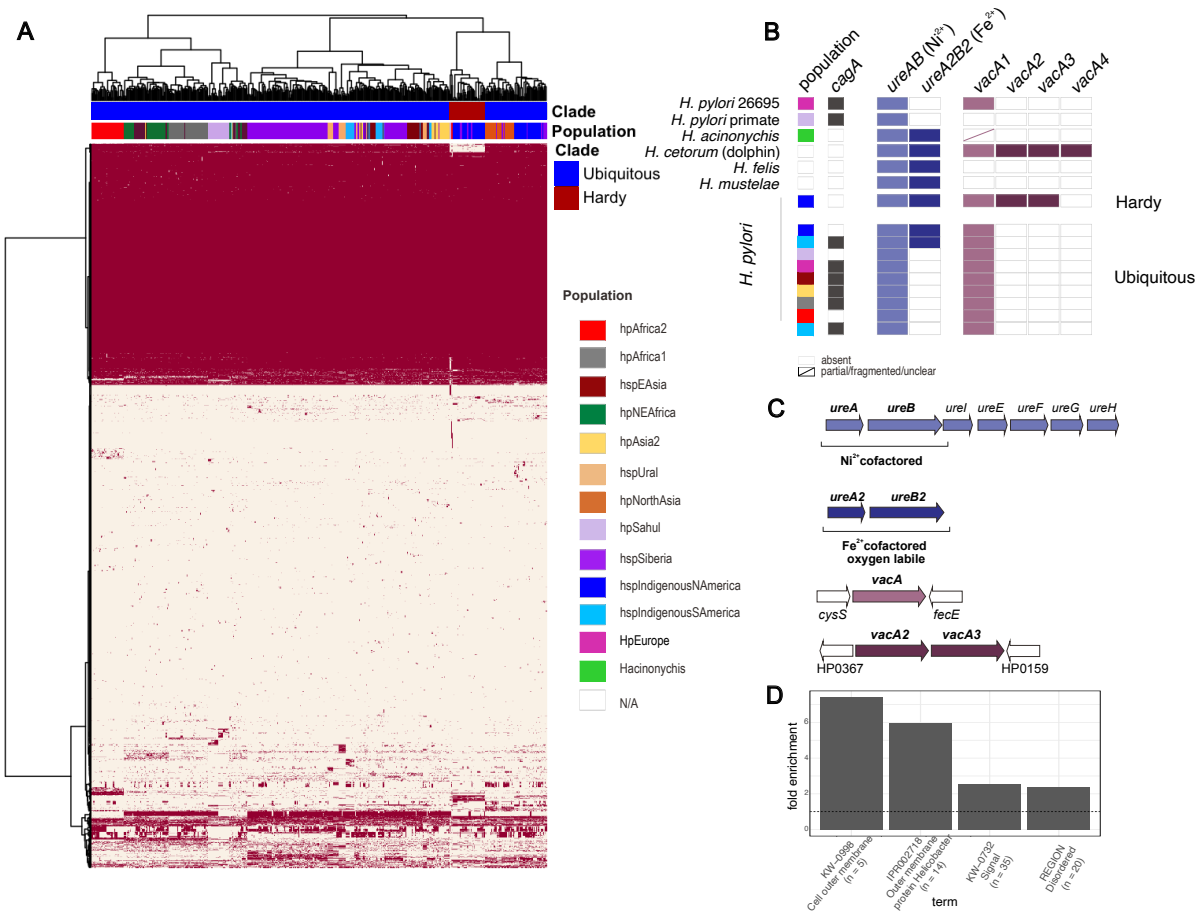


***Figure 4**: (A) Pangenome analysis on a sample of the global dataset: presence in red and absence in white of the different genes. The strains are colored based on their population and their ecospecies. (B) Presence and absence of cagA and the different types of vacA and ureA/B in a subset of the 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 (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 (*14*) this feature has otherwise been entirely associated with *Helicobacter* residing in obligate carnivore animal species: *H. 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 ecospecies.

Genes in molybdenum-related pathways (Mo uptake, molybdopterin biosynthesis) are absent in Hardy strains as previously reported for three of the strains (*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 ecospecies confirmed a striking enrichment of *Helicobacter* outer membrane proteins (21/100 genes, $p<0.05$, Figure 4D) (Table S4 for the list of differentiated genes and their categories, Table S5 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 ecospecies 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 (AlpA/HopB, and AlpB/HopC). Several of Hardy Hop-encoding genes were more closely related to *H. acinonychis* counterparts than to Ubiquitous H. pylori. Also, four OMP-coding genes are unique to the Hardy genomes from human hosts of which one is shared with H. acinonychis (exemplified by HopE in Figure S9D). One of these resemble BabA, which is missing in the Hardy strains, but major alterations in the cysteine-loop forming motifs involved in glycan binding suggest differences in binding specificity.

Two transcriptional regulators had fixed genetic differences in the Hardy ecospecies: 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 genes 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 ecospecies suggests these networks have been co-adapted to respond to distinct gastric environments.

**Discussion**

Our analysis is consistent with two distinct ecospecies of *H. pylori* emerging early in our evolution and both spreading around the world in our stomachs. Our definition of an ecospecies is a genetic strategy that is sufficiently distinct to have created its own independently evolving gene pool for a sizeable fraction of the genome. We have found Hardy strains from at least three different human *H. pylori* populations as well as *H. acinonychis*. The ecospecies arose long before the split of Khoisan populations from other extant human populations, extending the likely length of the association between *H. pylori* and man by at least 50%.

Previous authors have noticed unusual *H. pylori* strains which we now assign to the Hardy ecospecies. One set of authors sequenced a single genome (Aklavik86) from what they described as a “novel lineage”. Based on its distinctive gene content and 140 genome rearrangements in comparison with others strain from the same location, they speculated it might have arisen through a bacterium jumping from an animal reservoir to humans (*14*). Our analysis shows that the genome is a typical Hardy isolate. A scenario involving host jump to humans does not provide an obvious explanation for the wide geographical

distribution of both ecospecies 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 ecospecies (Figure S9) nor an obvious candidate for the reservoir species. Therefore, it is likely that the divergence and continuing evolution of the ecospecies has taken place within the ancestors of modern humans.

Another set of authors (*18*) noticed the similarity in gene content between strains isolated from Indigenous Americans 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, the strains are much more similar to other *H. pylori* than to *H. acinonychis*. However, this similarity reflects recent gene flow. Our analysis implies that *H. acinonychis* shares a common clonal ancestor with other strains from the Hardy ecospecies at around the time the Khoisan separated from other human populations. The most recent common ancestor of *any pair* of Hardy and Ubiquitous 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 (*14*). 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 ecospecies, 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 ecospecies has been maintained continuously by natural selection in each of the populations where the ecospecies co-occur.

Many bacterial species evolve distinct ecotypes, which are genetic strategies that give them a fitness advantage over other members of the same species within particular niches (*22*). If these strategies become more distinct from each other, so that intermediate genotypes become unfit, and their genetic basis becomes more complex, they can effectively suppress genetic exchange between ecotypes within a fraction of the genome. Within this fraction of the genome, the ecotypes can become good biological species and we refer to them as ecospecies, since genetic exchange is suppressed specifically by selection on ecological traits and does not entail full speciation. The Hardy ecospecies is sufficiently distinct with Ubiquitous *H. pylori* in the differentiated fraction of the genome (dS=0.34), that it would be considered a different species based on usual criteria applied to bacteria (*23*), if this divergence was found across the genome.

*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 (*24*). *V. parahaemolyticus* has a deeply diverged ecotype (EG1a), involving extensive coadaptation between loci encoding lateral flagella and genes involved in cell wall biosynthesis. This ecotype is also shared between geographical gene pools and has 18 diverged genome regions (*25*). Therefore, it also meets the ecospecies criteria, as well as less differentiated ecotypes that do not. Since both highly recombining species identified to

date have clearly defined ecospecies, it is an open question how frequent they are are within bacteria as a whole.

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 *H. pylori* ecospecies with different dietary capacities both dispersing widely throughout the globe and persisting together stably over long periods in several human populations. However, the Hardy ecospecies 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 (*26*) and therefore that the first human migrations mainly transmitted the meat-adapted Hardy ecospecies, while later migrations brought the Ubiquitous ecospecies, which is better adapted to a broader diet range, including omnivory. The only locations where the Hardy ecospecies 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 ecospecies has been isolated from human populations with significant gastric disease risk (*14*) and *H. acinonychis* causes severe gastritis and is a frequent cause of cheetah death in captivity (*27*). However, it remains to be seen if the Hardy ecospecies can also be defined as a "pathotype". Our hypothesis is that distinct ecospecies have evolved as an adaptation to human dietary variation, however a large fraction of the genes that are differentiated between ecospecies are in outer membrane proteins, including adhesins, 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 Helicobacter whole genome sequences from public and private sources, including 4254 *H. pylori* and 2 *H. acinonychis* genomes publicly available in Enterobase (*28*) (as of Jun 6$^{th}$ 2022) and 286 samples available in NCBI, BIGs and figshare; and 3635 *H. pylori* novel genomes from different geographic regions around the world. The novel sequences included 2126 isolates 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; a large number of worldwide DNA samples contributed by Prof Mark Achtman, of which 920 were sequenced either using the Illumina MiSeq platform at University of Gothenburg or at Novogene Co, Ltd, UK using the Illumina NOVA PE150 platform; and another 190 sequenced at University of Warwick, UK. Lastly, 266 strains from Iran, collected by Abbas Yadegar, and 133 genomes from different parts of the world, including 89 from the Swedish Kalixanda cohort (*29*), were all sequenced on the Illumina MiSeq platform at Karolinska Institutet and the University of Gothenburg, Sweden.

Samples from the Yamaoka lab were sequenced at Novogene Co., Ltd., Beijing, China with the Illumina NOVA PE150 platform and assembled using the SPAdes genome assembler v3.15.3 (*29*) by down sampling read depth to 100 bp, specifying a genome size of 1.6 Mbp and enabling the --careful option. Of the samples obtained from Prof. Achtman, 920 were sequenced either using the Illumina MiSeq platform at University of Gothenburg or at Novogene Co, Ltd, UK using the Illumina NOVA PE150 platform; and another 190 sequenced at University of Warwick, UK. Remaining sequences were sequenced on the Illumina MiSeq platform at Karolinska Institutet and the University of Gothenburg, Sweden and assembled using the BACTpipe pipeline (https://doi.org/10.5281/zenodo.4742358).
We then filtered out redundant genomes, 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 (*30*), to obtain a final total of 6626 genomes, correspond to 2663 genomes first reported in this work and 2965 publicly available strains (Table S1).

**Core Genome variant calling**
Single nucleotide polimorphisms from the core genome (core SNPs) were called using a MUMmer v3.20 (*31*)-based algorithm as previously described (*32*). 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, 33, 34*) 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) (*35*).

**Phylogenetic analysis**
The different Maximum Likelihood (ML) phylogenetic trees were obtained with FastTree (v2.1.10, option -nt) (*36*) , 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) (*37*).

**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 ecospecies 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 ecospecies. The ubiquitous strains were coded as 0 (171 strains) while Hardy strains were coded as 1 (38 strains). The GWAS was performed using the R (*38*) package bugwas (v0.0.0.9000) (*39*), 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, 3942 out of 278,019 core biallelic coding SNPs were significantly associated with the Hardy clade.

To reinforce the results obtained by the GWAS, we calculated the per-site $F_{ST}$ between the Ubiquitous and Hardy ecospecies via the R package PopGenome (*40*) using the same set of strains and SNPs. We considered a SNP to be differentiated between the Hardy and Ubiquitous ecospecies if it was significantly associated with the ecospecies by the GWAS ($-\log_{10}(p) > 10$) and highly differentiated between the two groups based on its $F_{ST}$ value ($F_{ST} > 0.9$). Of the core biallelic SNPs, we found 2595 differentiated coding SNPs and 175 differentiated intergenic SNPs. A SNP was considered to be undifferentiated if its $-\log_{10}(p)$ was less than 10 and its $F_{ST}$ less than 0.5 (257,780 coding and 8964 intergenic SNPs). All the other SNPs (7,773 CDS and 607 intergenic) were considered to be intermediate.
Following the separation of the SNPs into three classes, we also distinguished three types of genes, based on the genes present in the 26695 reference genome: differentiated (100 genes, Table S2), containing at least five differentiated SNPs; undifferentiated (1241 genes), with only undifferentiated SNPs, and the remaining genes (236 genes) which we 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 (*41*), using the proteome of the 26695 reference strain as reference. Then, .gff files were used as input for Panaroo's v1.2.8 (*42*) 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 (*43*) 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.

**Genome structure comparison**
Significant structural variations like inversions, gaps, repeats, and gene cluster rearrangement can be easily visualized using a dot plot. To investigate the genome structure similarity and the difference between Hardy and Ubiquitous groups, we used the Gepard program (*44*) (v1.40 jar file from https://cube.univie.ac.at/gepard) to make the dot plot. Different comparisons were considered: Hardy vs. Hardy, Hardy vs. Ubiquitous, and Ubiquitous vs. Ubiquitous, and against *H. cetorum*. Publicly available genome sequences were downloaded from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The Gepard internal DNA substitution matrix (edna.mat) was selected to generate the alignment and plot. The lower color limit was 50 to reduce noise and emphasize significant regions. Window size and word length were default values of 0 and 10, respectively.

**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 (*45*). Three hypothetical proteins were removed, as they lacked a unique ID. The background set of genes chosen (Table S4 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 (*46*) (*47*). 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. S. Suerbaum, P. Michetti, Helicobacter pylori infection. *N Engl J Med* **347**, 1175-1186 (2002).
2. M. Amieva, R. M. Peek, Jr., Pathobiology of Helicobacter pylori-Induced Gastric Cancer. *Gastroenterology* **150**, 64-78 (2016).
3. P. Olbermann *et al.*, A global overview of the genetic and functional diversity in the Helicobacter pylori cag pathogenicity island. *PLoS Genet* **6**, e1001069 (2010).
4. J. Parsonnet, G. D. Friedman, N. Orentreich, H. Vogelman, Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection. *Gut* **40**, 297-301 (1997).
5. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
6. D. Falush *et al.*, Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582-1585 (2003).
7. Y. Moodley *et al.*, Age of the association between Helicobacter pylori and man. *PLoS Pathog* **8**, e1002693 (2012).
8. Y. Moodley *et al.*, Helicobacter pylori's historical journey through Siberia and the Americas. *Proc Natl Acad Sci U S A* **118**, (2021).
9. S. Fan *et al.*, Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923-939 e914 (2023).
10. N. Hollfelder, G. Breton, P. Sjodin, M. Jakobsson, The deep population history in Africa. *Hum Mol Genet* **30**, R2-R10 (2021).
11. M. Eppinger *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. L. Kennemann *et al.*, Helicobacter pylori genome evolution during human infection. *Proc Natl Acad Sci U S A* **108**, 5033-5038 (2011).
13. D. Kersulyte, M. Rossi, D. E. Berg, Sequence divergence and conservation in genomes of Helicobacter cetorum strains from a dolphin and a whale. *PLoS One* **8**, e83177 (2013).
14. D. Kersulyte *et al.*, Complete Genome Sequences of Two Helicobacter pylori Strains from a Canadian Arctic Aboriginal Community. *Genome Announc* **3**, (2015).
15. E. L. Carter, D. E. Tronrud, S. R. Taber, P. A. Karplus, R. P. Hausinger, Iron-containing urease in a pathogenic bacterium. *Proc Natl Acad Sci U S A* **108**, 13095-13099 (2011).
16. J. Stoof *et al.*, Inverse nickel-responsive regulation of two urease enzymes in the gastric pathogen Helicobacter mustelae. *Environ Microbiol* **10**, 2586-2597 (2008).
17. A. H. van Vliet *et al.*, Differential regulation of amidase- and formamidase-mediated ammonia production by the Helicobacter pylori fur repressor. *J Biol Chem* **278**, 9052-9057 (2003).
18. H. Gressmann *et al.*, Gain and loss of multiple genes during the evolution of Helicobacter pylori. *PLoS Genet* **1**, e43 (2005).
19. M. Kawai *et al.*, Evolution in an oncogenic bacterial species with extreme genome plasticity: Helicobacter pylori East Asian genomes. *BMC Microbiol* **11**, 104 (2011).
20. J. T. Loh *et al.*, Delineation of the pH-Responsive Regulon Controlled by the Helicobacter pylori ArsRS Two-Component System. *Infect Immun* **89**, (2021).
21. K. P. Haley, J. A. Gaddy, Metalloregulation of Helicobacter pylori physiology and pathogenesis. *Front Microbiol* **6**, 911 (2015).

22. F. M. Cohan, Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla. *Microbiol Spectr* **5**, (2017).
23. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**, 1929-1940 (2006).
24. C. Yang, Y. Cui, X. Didelot, R. Yang, D. Falush, Why panmictic bacteria are rare. *Bioarxiv*, (2019).
25. Y. Cui *et al.*, The landscape of coadaptation in Vibrio parahaemolyticus. *Elife* **9**, (2020).
26. J. Dembitzer, R. Barkai, M. Ben-Dor, S. Meiri, Levantine overkill: 1.5 million years of hunting down the body size distribution. *Quaternary Science Reviews* **276**, 107316 (2022).
27. J. G. Kusters, A. H. van Vliet, E. J. Kuipers, Pathogenesis of Helicobacter pylori infection. *Clin Microbiol Rev* **19**, 449-490 (2006).
28. Z. Zhou *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).
29. P. Aro *et al.*, Peptic ulcer disease in a general adult population: the Kalixanda study: a random population-based study. *Am J Epidemiol* **163**, 1025-1034 (2006).
30. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).
31. S. Kurtz *et al.*, Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).
32. Y. Cui *et al.*, Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Mol Biol Evol* **32**, 1396-1410 (2015).
33. B. Linz *et al.*, An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915-918 (2007).
34. Y. Moodley, B. Linz, Helicobacter pylori Sequences Reflect Past Human Migrations. *Genome Dyn* **6**, 62-74 (2009).
35. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
36. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
37. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
38. R. C. Team, R: A language and environment for statistical computing. (2021).
39. S. G. Earle *et al.*, Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041 (2016).
40. B. Pfeifer, U. Wittelsburger, S. E. Ramos-Onsins, M. J. Lercher, PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* **31**, 1929-1936 (2014).
41. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
42. G. Tonkin-Hill *et al.*, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
43. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

44. J. Krumsiek, R. Arnold, T. Rattei, Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026-1028 (2007).
45. W. Huang da, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
46. Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000).
47. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (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, Tam Anh General Hospital, Ho Chi Minh city, 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)

Jukka Ronkainen, MD, PhD, Center for Life Course Health Research, University of Oulu; Finland and Primary Health Care Center, Tornio, Finland (jukka.ronkainen@fimnet.fi)

Pertti Aro, MD, PhD, Arokero Oy, Tornio, Finland (pertti.aro@fimnet.fi)

Lars Agreus, MD, PhD, Division of Family Medicine, Karolinska Institutet, Stockholm, Sweden (lars.agreus@ki.se)

Anna Andreasson, PhD, Stress Research Institute, Department of Psychology, Faculty of Social Sciences, Stockholm University, Sweden (anna.andreasson@su.se)

Lars Engstrand, MD, PhD, Center for Translational Microbiome Research, Department for Microbiology, Tumor, and Cell Biology, Karolinska Institutet, Stockholm, Sweden (lars.engstrand@ki.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. Sequencing and database development was also funded in part by Wellcome Trust grant 202792/Z/16/Z to Mark Achtman. 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.