Article

An ancient ecospecies of *Helicobacter pylori*

https://doi.org/10.1038/s41586-024-07991-z

Received: 4 August 2023

Accepted: 23 August 2024

Published online: 16 October 2024

Open access

Check for updates

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], Helicobacter Genomics Consortium*, Difei Wang[6], Abbas Yadegar[7], Lisa M. Olsson[8], Zhemin Zhou[9], Yoshio Yamaoka[2,3,10,11] ✉, Kaisa Thorell[12] ✉ & Daniel Falush[1] ✉

Helicobacter pylori disturbs the stomach lining during long-term colonization of its human host, with sequelae including ulcers and gastric cancer[1,2]. Numerous *H. pylori* virulence factors have been identified, showing extensive geographic variation[1]. Here we identify a 'Hardy' ecospecies of *H. pylori* that shares the ancestry of 'Ubiquitous' *H. pylori* from the same region in most of the genome but has nearly fixed single-nucleotide polymorphism differences in 100 genes, many of which encode outer membrane proteins and host interaction factors. Most Hardy strains have a second urease, which uses iron as a cofactor rather than nickel[3], and two additional copies of the vacuolating cytotoxin VacA. Hardy strains currently have a limited distribution, including in Indigenous populations in Siberia and the Americas and in lineages that have jumped from humans to other mammals. Analysis of polymorphism data implies that Hardy and Ubiquitous coexisted in the stomachs of modern humans since before we left Africa and that both were dispersed around the world by our migrations. Our results also show that highly distinct adaptive strategies can arise and be maintained stably within bacterial populations, even in the presence of continuous genetic exchange between strains.

To characterize the global diversity of this important human pathogen, we assembled a dataset of 9,188 *Helicobacter* genome sequences (comprising 9,186 *H. pylori* and two *Helicobacter acinonychis*) from humans and other hosts around the world, including from many undersampled populations (Supplementary Tables 1 and 2). Following quality control (see Methods) our subsequent analysis was performed on 6,864 *H. pylori* genomes and two *H. acinonychis*. *H. acinonychis* has been isolated from big cats in zoos and represents a human-to-animal host jump[4]. *H. pylori* has also been occasionally isolated from animals, including four from primates at the UC Davis Primate Research Center, which has housed rhesus macaques and cynomolgus monkeys. In addition to published genomes[5–31], the dataset contains 2,916 unpublished genomes from 56 countries following quality control, including a large new set of samples from Southeast Asia and Iran, as well as strains previously defined only by multilocus sequence typing, among them a large number of genomes from Siberia. Following previous practice[24,25], chromosome painting was used to assign the strains to 13 populations (designated hp) and less differentiated subpopulations (designated hsp), each of which have different geographic distributions (Supplementary Table 2 and Extended Data Fig. 1).

'Hardy' and 'Ubiquitous' strains

We noticed that we got surprisingly different answers for how strains clustered, depending on the method used (Extended Data Figs. 2a, 3 and 4). A subset of 48 Hardy strains from Chile, Siberia, Canada and the United States, as well as two strains from the UC Davis Primate Center, formed a clade in phylogenetic trees (Extended Data Fig. 2a). The same strains were separated from the Ubiquitous others on a principal components analysis (PCA) plot (Extended Data Fig. 3). Hardy strains were named as such because they were isolated from individuals living in locations that most of the world's population would consider physically inhospitable, whereas Ubiquitous strains are found in humans everywhere, including the same areas as Hardy. The human Hardy strains were assigned to three different populations, hspSiberia, hspIndigenousNAmerica and hspIndigenousSAmerica, and fineSTRUCTURE[32] grouped them with strains from their own geographic regions (Extended Data Fig. 4). This unexpected pattern led us to investigate the origin and evolution of these distinct groups, including prediction of the functional consequences of their genomic differences.

To understand why strains from different locations and ancestry profiles grouped together by phylogenetic analysis and PCA,

[1]Shanghai Institute of Immunity and Infection, Chinese Academy of Sciences, Shanghai, China. [2]Department of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Yufu, Japan. [3]Universitas Airlangga, Surabaya, Indonesia. [4]Gastroenterology Unit, Department of Medicine and Center of Excellence in Digestive Diseases, Thammasat University, Bangkok, Thailand. [5]Cho Ray Hospital, Ho Chi Minh City, Vietnam. [6]Cancer Genomics Research Lab, Frederick National Lab for Cancer Research, Rockville, MD, USA. [7]Foodborne and Waterborne Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [8]The Wallenberg Laboratory, Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [9]Pasteurien College, Suzhou Medical College, Soochow University, Suzhou, China. [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, Gothenburg, Sweden. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: yyamaoka@oita-u.ac.jp; kaisa.thorell@gu.se; daniel.falush@ips.ac.cn

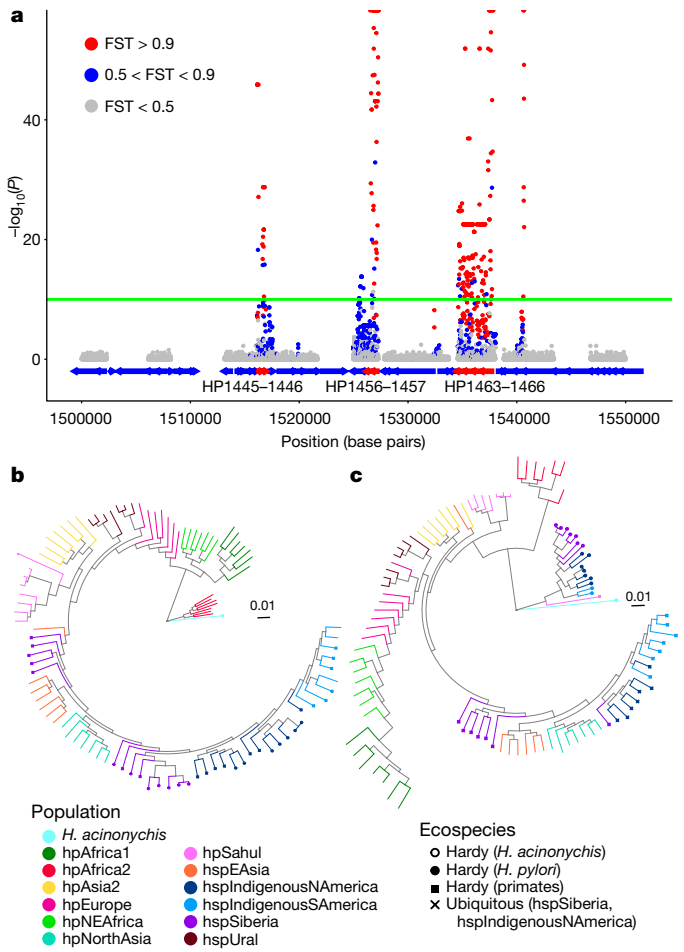


Fig. 1 | Differentiation between Hardy and Ubiquitous strains is localized in the genome. a, Manhattan plot from GWAS analysis of Hardy versus Ubiquitous strains from hspSiberia and hspIndigenousNAmerica (zoomed in between 1.50 and 1.55 megabase pairs; full plot shown in Extended Data Fig. 5). Genes are indicated by blue and red (differentiated genes) arrows. Green line indicates significance threshold ($-\log_{10}(P) = 10$, which is based on a Bayesian Wald test with a Bonferroni correction for multiple testing, using a significance threshold before correction of 3×10^{-5}, and 285,792 tested SNPs). Points are coloured based on FST (fixation index) values; half-points at the top of the plot indicate estimated $P = 0$ and FST = 1. HP1445 and so on are *H. pylori* genes based on the annotation of the 26695 strain. **b**,**c**, Phylogenetic trees for undifferentiated (**b**) and differentiated (**c**) genes from a representative subset of strains (see Extended Data Fig. 2b,c for trees of the whole dataset). Branches are coloured based on population. Strains from the Hardy clade are indicated by a filled circle at the end of the branch.

we performed genome-wide analysis (genome-wide association study, GWAS) 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 (Fig. 1a and Extended Data Fig. 5). We therefore split the genome alignment into genes with nearly fixed differences between the Hardy and Ubiquitous clades (100 of 1,577), genes with no differentiated single-nucleotide polymorphisms (SNPs) (1,034) and intermediate genes that did not fall into the other two categories (443) to characterize patterns of genetic differentiation separately for these fractions of the genome (Supplementary Table 3).

Hardy strains have entirely different genetic relationships with other *H. pylori* at differentiated versus undifferentiated genes, based on phylogenetic trees (Fig. 1b,c and Extended Data Fig. 2b,c). The tree of undifferentiated genes (Fig. 1b and Extended Data Fig. 2b) is consistent with the evolutionary relationships within *H. pylori* established in previous analyses[15,33,34], with Hardy strains from hspIndigenousSAmerica, hspSiberia and hspIndigenousNAmerica populations clustering with Ubiquitous strains isolated from the same locations. The longest branch in the tree separates strains from the hpAfrica2 population and *H. acinonychis* from other *H. pylori*. Concordant with previous inferences[4,34], the tree is rooted on this branch. hpAfrica2 strains originated in Khoisan populations in southern Africa[34], who are the oldest-branching group in the human population tree[35]. By contrast, at differentiated genes (Fig. 1c and Extended Data Fig. 2c), Hardy strains isolated from humans are the most genetically distinct *H. pylori*, branching more deeply than hpAfrica2, and are more closely related to *H. acinonychis*. This divergence causes the Hardy strains to branch separately from Ubiquitous strains from the same population in phylogenetic analyses based on the whole genome (Extended Data Fig. 2a).

The two primate Hardy isolates are assigned to the hpSahul population, found in humans in Indonesia, Papua New Guinea and Australia, and cluster with other isolates from this population in the tree of undifferentiated genes (Fig. 1b and Extended Data Fig. 2b). However, the ancestry profile of these isolates is clearly distinct from that of any other hpSahul isolate in the database (Extended Data Fig. 6), suggesting that the host jump may have been ancient. Two further strains from the UC Davis Primate Research Center belong to the hpAsia2 population (Supplementary Table 2) and appear to be typical Ubiquitous *H. pylori* (data not shown).

Origin of ecospecies

To understand the global spread of Hardy and Ubiquitous, we next investigated their clonal relationships. Bacteria reproduce by binary fission, meaning that there is a single genealogical tree representing the clonal (cellular) relationships of any sample, but in *H. pylori* homologous recombination is frequent, transferring DNA between strains that coinfect the same host[36]. Recombination of short tracts is unlikely to affect genome order, especially for large rearrangements. Therefore, rearrangements can potentially provide a marker of clonal descent, even in the presence of high recombination rates. Dot plots showed many more rearrangements between any pair of Hardy and Ubiquitous strains than between strains of the same ecospecies (Extended Data Fig. 7), with around 140 rearrangements between two hspIndigenousNAmerica strains from the same ecospecies, compared with around ten within ecospecies[3]. These results show that, at the cellular level, the split between Hardy and Ubiquitous strains was ancient. Rearrangements disrupt homology between strains in the vicinity of breakpoints, but high rates of genetic exchange have continued between ecospecies, probably because of the short length of many fragments incorporated by homologous recombination in *H. pylori*[36]. Nevertheless, at the differentiated genes, Hardy and Ubiquitous strains have retained their distinct identities (Fig. 1c).

To reconcile the discordant evolutionary history of differentiated and undifferentiated genes and the ancient split between Hardy and Ubiquitous clonal backgrounds, we hypothesize that *H. pylori* split into two ecospecies before the split of hpAfrica2 from other populations (Fig. 2a). We define ecospecies as bacteria within the same species that have undergone species-level differentiation within a specific fraction of the genome, but otherwise having a single recombining gene pool. It is likely that differentiation in that fraction is maintained by strong ecologically mediated selection against intermediate genotypes. However, other forces, such as genetic mechanisms greatly restricting gene flow in parts of the genome, might also play a role.

We propose that both ecospecies were spread around the world by human migrations. Because the gene pools at differentiated loci are distinct, with each ecospecies having their own segregating SNPs, genetic relationships at these genes represent the separate histories of spread. The DNA may have been carried to their present locations

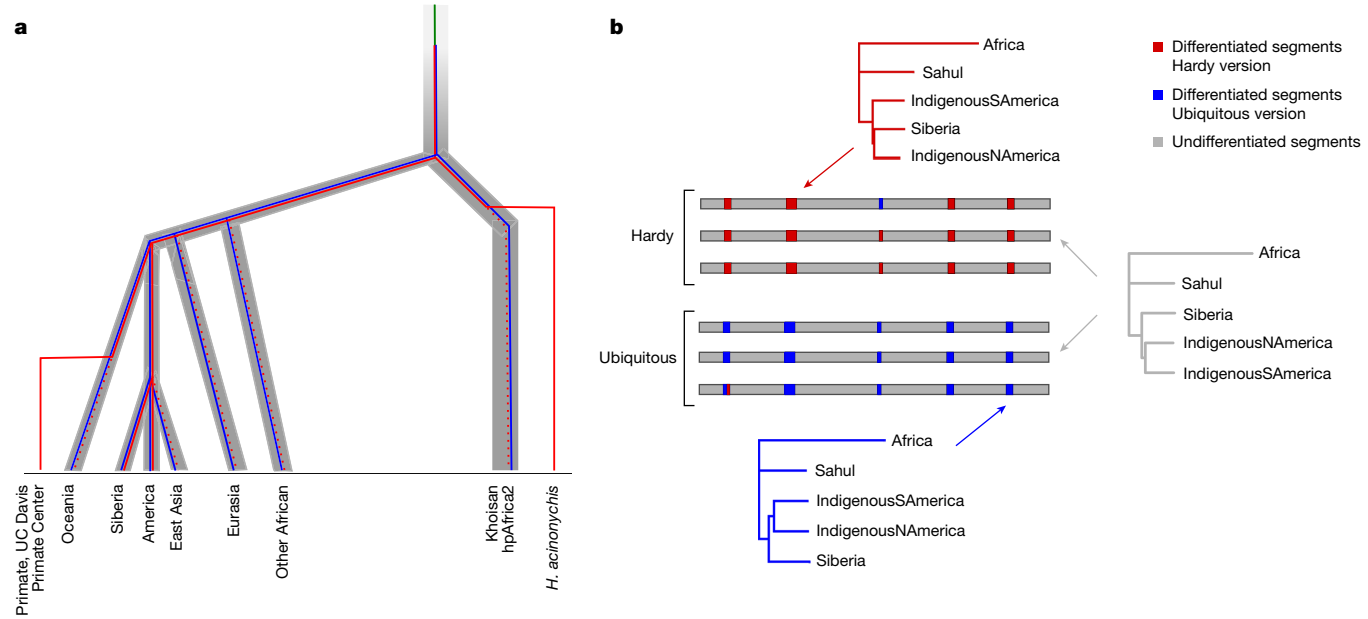


Fig. 2 | Divergence and spread of Hardy and Ubiquitous gene pools. **a**, Hypothesized scenario for the differentiation of *H. pylori* into two ecospecies and subsequent global spread (not to scale). Thick grey tree represents a simplified history of human population differentiation (based on ref. 49). *Helicobacter* evolution is represented by thinner lines, which are within the grey tree during periods of evolution in humans. Green line represents *H. pylori* before the evolution of the two ecospecies; blue line represents Ubiquitous ecospecies; red line represents Hardy ecospecies; dotted red lines indicate that Hardy strains have not yet been detected on the branch and therefore may have gone extinct. **b**, Phylogenetic population trees for differentiated (Hardy in red and Ubiquitous in blue) and undifferentiated (in grey) regions of the genomes. Trees were constructed considering only populations with both Hardy and Ubiquitous representatives. Pairwise distances between strains were calculated for the relevant genome regions, and population distances were calculated by averaging over pairwise strain distances. For Africa, we used *H. acinonychis* strains for the Hardy differentiated gene trees and hpAfrica2 strains for the Ubiquitous differentiated gene tree. For the undifferentiated gene tree, we averaged over the relevant Hardy and Ubiquitous strains.

by different human migrations whereas undifferentiated genes reflect an amalgam of both histories. Phylogenetic trees constructed for Hardy and Ubiquitous strains from the same populations at differentiated genes show similar, parallel, histories (Fig. 2b), albeit with minor differences. Specifically, for Ubiquitous and undifferentiated segments, hspIndigenousSAmerica and hspIndigenousNAmerica strains cluster together. In Hardy strains, hspIndigenousSAmerica strains cluster in a deeper position in the tree, hinting that they may have been brought to the Americas by a distinct, and probably earlier, human migration.

The ratio of non-synonymous to synonymous diversity (dN/dS) indicates the degree of functional constraint on genes, and is also impacted by positive selection (Extended Data Fig. 8a,b). Within Ubiquitous *H. pylori*, dN/dS levels are lower in differentiated genes than in undifferentiated (0.125 versus 0.171), implying that differentiated genes are on average more functionally constrained than the remainder of the genome. However, at differentiated genes, dN/dS levels are elevated in pairwise comparisons involving Hardy and hpAfrica2 strains compared with those in Ubiquitous strains (0.178 versus 0.125). This difference implies that, following differentiation between the two ecospecies, there was an elevated rate of evolution at functional sites at these genes, which is probably associated with specialization to distinct niches.

The fineSTRUCTURE results, with similar levels of coancestry within and between ecospecies in particular geographical locations, demonstrate that there is frequent recombination between ecospecies in most of the genome (Extended Data Fig. 4). The question is, then: what force has maintained the differentiated regions? If this force is purifying selection to retain ecospecies-specific function, 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 purifying selection (Fig. 3a). For many of the differentiated genes there are Hardy strains that have Ubiquitous haplotypes, and vice versa, but each of these haplotypes is present in only one or a small number of strains.

Further evidence that selection has maintained 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—that is, the predominant allele in the Hardy clade—as a function of genetic distance to representatives of the Ubiquitous ecospecies from populations in which Hardy ecospecies strains were isolated (Fig. 3b). Distantly related strains (for example, 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 Fig. 3b. There are 11 strains (Supplementary Table 2) that are clear outliers from the trend line in Fig. 3b, and all of these derive from geographic regions in which the Hardy ecospecies is found (Fig. 3c). The same outliers appear in a plot of the number of differentiated blocks against the number of differentiated SNPs (Extended Data Fig. 9). Because genetic exchange tends to progressively homogenize ancestry in the absence of natural selection, the absence of outlier strains elsewhere suggests that Hardy alleles are deleterious on a Ubiquitous genetic background and that the outliers themselves can be explained by recent, local genetic exchange. Furthermore, the absence of outliers elsewhere implies that Hardy itself has had a limited geographical range in recent times.

Our hypothesis presented in Fig. 2a implies that Hardy progressively diverged from Ubiquitous by new mutations in the differentiated genes. To investigate evidence for alternative scenarios in which acquisition of DNA from other *Helicobacter* played a role, we assembled a panel of *Helicobacter* species from other mammals, excluding *H. acinonychis* (Supplementary Table 4). We performed BLAST on each of the 100 differentiated genes to identify close matches. In total, 49 genes had significant BLAST matches (Supplementary Table 3), the closest of which was always in *Helicobacter cetorum*, whose hosts are dolphins

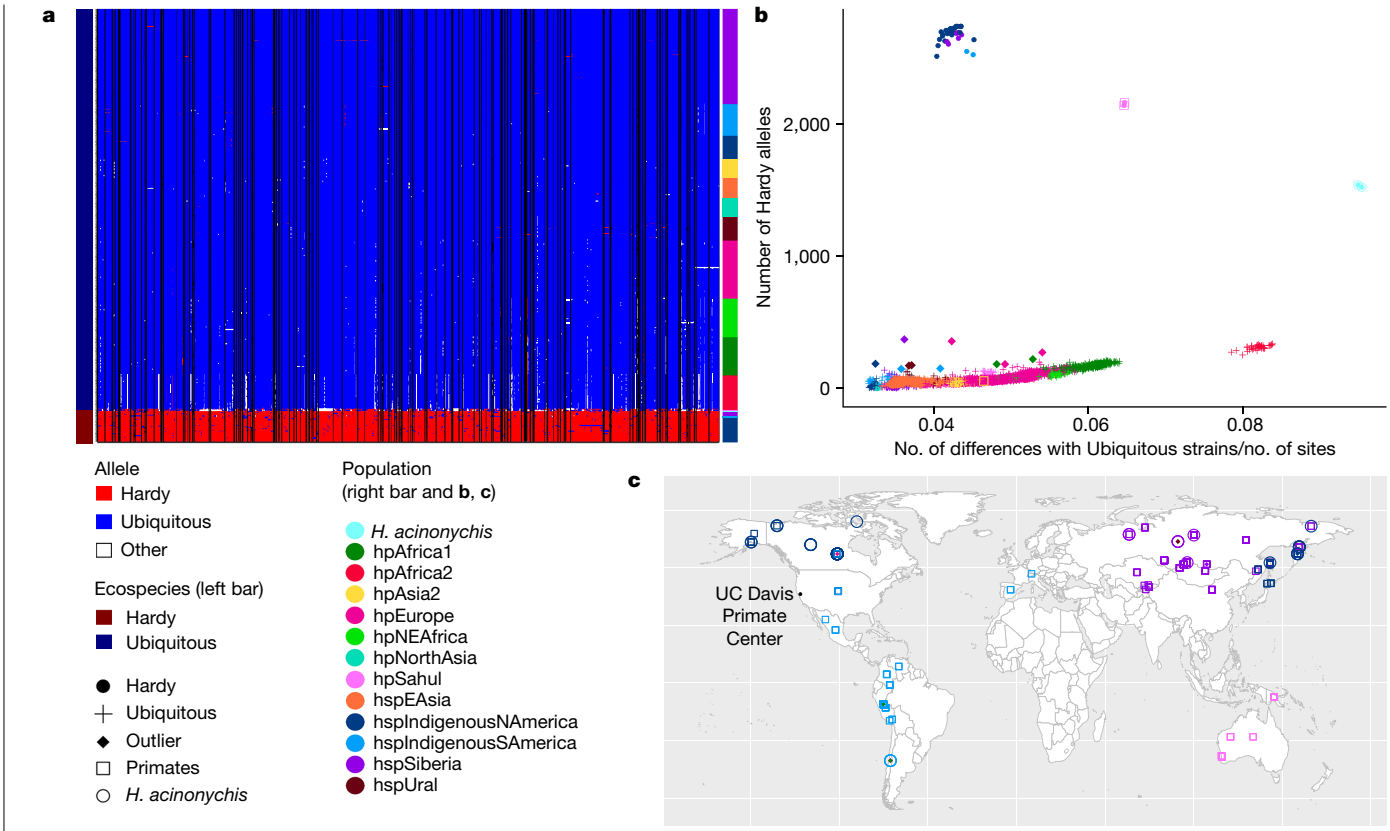


Fig. 3 | Geographic distribution of Hardy and Ubiquitous haplotypes. **a**, Haplotypes at the SNPs that are differentiated between the two ecospecies for hspSiberia and hspIndigenousNAmerica strains, and randomly selected representatives from other populations. The major Hardy allele is represented in red and the major Ubiquitous allele in blue; other alleles are shown in white. Black vertical lines separate different genes. **b**, Number of Hardy alleles as a function of average genetic distance with Ubiquitous ecospecies strains from hspIndigenousSAmerica, hspSiberia and hspIndigenousNAmerica. Dots are coloured based on their populations with Ubiquitous strains represented by crosses and Hardy strains represented by circles. Filled diamonds indicate Ubiquitous outlier strains with a higher number of Hardy alleles than expected for a strain at that genetic distance. **c**, Map showing location of Hardy (circles) and Ubiquitous (squares) strains from hspIndigenousSAmerica, hspSiberia, hspIndigenousNAmerica and hpSahul populations, as well as strains that were outliers (filled diamonds) compared with their population in the number of Hardy alleles they have. For some strains, information on where they were isolated (latitude and longitude) was missing, in which case we entered the coordinates of their country of isolation.

and whales, and is the closest known relative of *H. pylori* other than *H. acinonychis*.

The results of the BLAST analysis show no evidence for importation of DNA to *H. pylori* from other *Helicobacter* spp. For 29 of the genes, the evolutionary distances, visualized using phylogenetic trees (Supplementary Fig. 1a), were consistent with the scenario shown in Fig. 2a, with Hardy and Ubiquitous versions diverging from each other before the split of hpAfrica2 and *H. acinonychis* from other populations. For 13 of the genes, the tree instead implied that the Hardy and Ubiquitous versions of the gene diverged from each other following the split between hpAfrica2 and other human *H. pylori*, with hpAfrica2 and *H. acinonychis* versions of the gene clustering together (Supplementary Fig. 1b). Three genes (*horL*, hp0599, encoding methyl-accepting chemotaxis protein and *lp220*) showed signs of gene flow between Hardy and Ubiquitous versions (Supplementary Fig. 1c). Three genes, including *vacA* (discussed below), had a history involving gene duplication (Supplementary Fig. 1d) whereas, for the gene encoding the outer membrane protein HopF, the Hardy version diverged from the Ubiquitous version before the split with *H. cetorum* (Supplementary Fig. 1e). 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 in which the pattern of polymorphism was suggestive of import of genetic material from other *Helicobacter* species.

Ecospecies function

Clues to the ecological basis of ecospecies divergence are provided by the functions of differentiated and accessory genes from the Hardy ecospecies (Supplementary Table 3). Hardy strains, including *H. acinonychis*, form their own clade in a tree constructed based on accessory genome analysis (Fig. 4a). Only five (the two primate strains included) out of 48 Hardy strains have an intact *cag* pathogenicity island, which is a much lower rate than for the Ubiquitous ecospecies (472 of 673 strains) but sufficient to imply that the island can function and be maintained by natural selection on both ecospecies backgrounds (Fig. 4a,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 show an additional tandem pair of duplicated *vacA* genes (Fig. 4b,c and Supplementary Figs. 1 and 2).

Hardy strains carry an iron-cofactored urease (UreA2B2) in addition to the canonical nickel-dependent enzyme of *H. pylori* (Fig. 4b,c). Iron-cofactored urease is thought to facilitate survival in the low-nickel gastric environment of carnivores[37]. Although previously noted in a single Hardy *H. pylori* strain from northern Canada[3], 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)[38]. Pangenome analysis of non-*pylori Helicobacter* species isolated from animals (Fig. 4b and Supplementary Table 4) confirmed these associations and identified

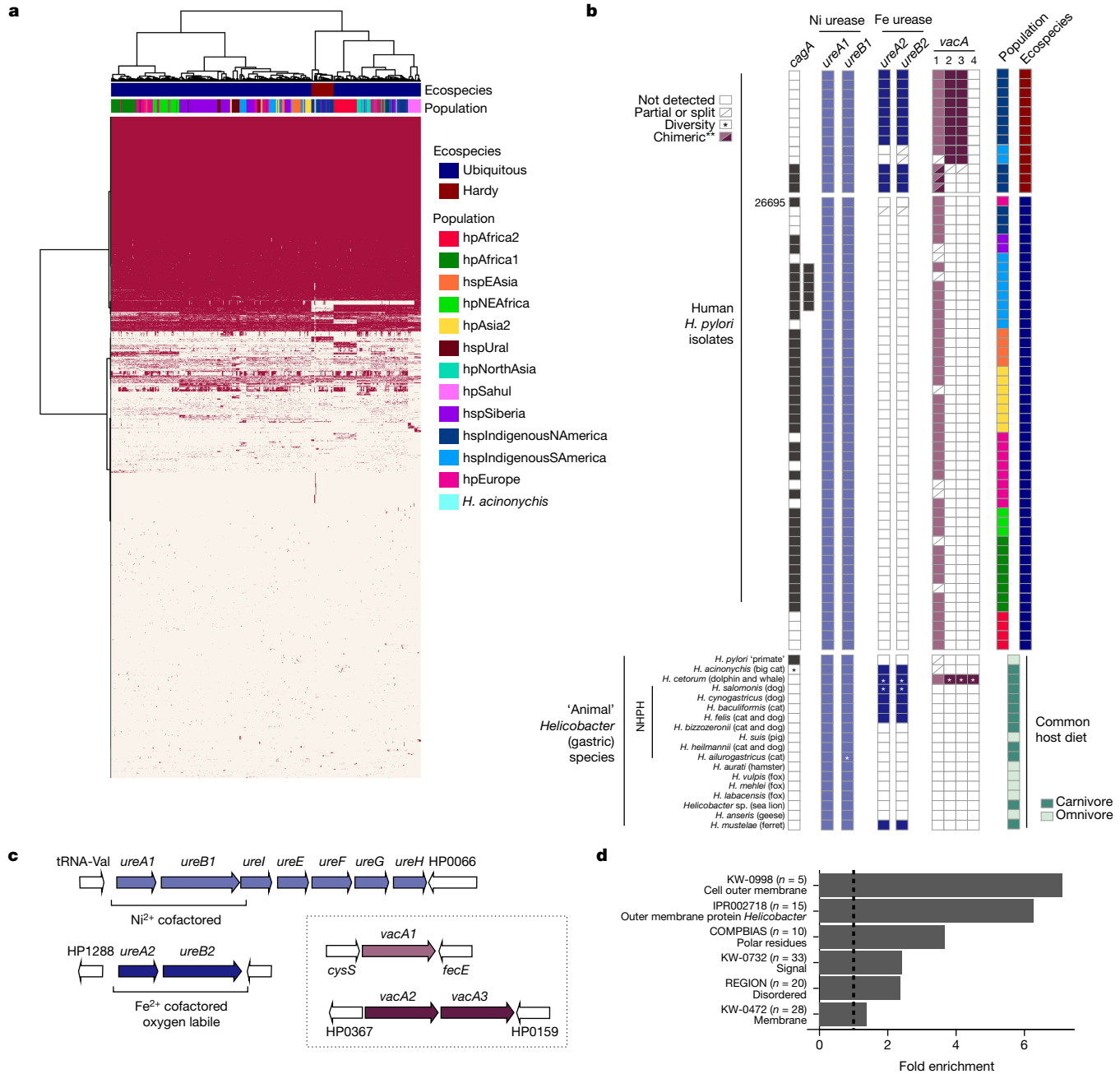


Fig. 4 | Genome composition in human and animal *Helicobacter*. **a**, Hierarchical clustering based on pangenome presence (red)/absence (white) for a sample of the global dataset. Strains are coloured based on their population and ecospecies. **b**, Top, presence/absence of *cagA* and different *vacA* and *ureA/B* types in Hardy and Ubiquitous *H. pylori* from humans. Included is a subset of complete genomes, two *cagA*$^+$ Hardy strains and reference strain 26695. Bottom, presence/absence in gastric *Helicobacter* isolated from animals. Common host(s) and their diets are indicated (Supplementary Table 4). **c**, Representative configuration and genomic context of *ureA/B* and *vacA* in Hardy *H. pylori* genomes, based on strain HpGP-CAN-006. Lighter-coloured arrows indicate genes present in both ecospecies, based on sequence and genomic context/synteny; darker-coloured arrows indicate Hardy-specific versions. **d**, Fold enrichment for significant ($P < 0.05$, one-sided Fisher's test, Benjamini correction) functional terms. Chimeric, potential chimeric version (Hardy + Ubiquitous); Diversity, differential presence/absence in analysed genomes within the species; NHPH. non-*H. pylori Helicobacter* spp.; tRNA, transfer RNA.

ureA2B2 homologues in additional species associated with carnivores (cats, dogs), including *H. cynogastricus*, *H. salomonis* and *H. baculoformis*. We did not detect *ureA2B2* in some gastric *Helicobacter* spp. associated with cats and dogs (for example, *H. heilmannii*, *H. bizzozeronii*); however, these species are known to cocolonize the stomach with other species that carry iron urease such as *H. felis*[39]. We did not detect *ureA2B2* in any species that colonize omnivores, in either the Hardy primate strains that also probably encounter a varied diet (Fig. 4c) or any enterohepatic species (data not shown). The *ureA2B2* locus reflects a very old duplication event, with versions of both genes also found in *H. cetorum* (Supplementary Figs. 3 and 4). Therefore, as we also observed for *vacA*, it is likely that the ancestral population had two versions of the gene but that one has been lost by Ubiquitous strains following evolution of the ecospecies.

Functional enrichment analysis of the 100 differentiated genes in the Hardy ecospecies confirmed a marked enrichment of *Helicobacter* outer membrane proteins (21 of 100 genes, $P < 0.05$; Fig. 4d and Supplementary Tables 3 and 5), and the list also included many

genes involved in cell envelope biogenesis. Four outer membrane protein-coding genes are unique to the Hardy genomes. One gene resembles the carbohydrate-binding adhesin BabA, which is missing in Hardy strains, but major alterations in the cysteine loop-forming motifs involved in glycan binding suggest differences in binding specificity.

Discussion

Our analysis has found that Indigenous people in Siberia and North and South America are infected by two distinct types of *H. pylori* that we have named Hardy and Ubiquitous. These types are distinguished by nearly fixed differences in 100 out of 1,577 genes (based on an alignment with the 26695 reference genome) and important differences in gene content, despite evidence of extensive genetic exchange throughout the genome. We designate these types as ecospecies, because within the differentiated fraction of the genome they have their own gene pools that are sufficiently different (minimum dS of 0.20 between Hardy and Ubiquitous strains and maximum average nucleotide identity (ANI) of 0.94; Extended Data Fig. 8e) that they would be considered a different species based on the usual criteria applied to bacteria of ANI < 0.95 (ref. 40), if this divergence was found across the genome. However, they are not species, because most of the genome is undifferentiated and there is no evidence that distinct types ever existed in the undifferentiated fraction of the genome. A similar pattern of genetic variation has been observed in *Vibrio parahaemolyticus*, in which the the EG1a genotype is differentiated at 26 genes in 18 genome regions[41], despite high recombination elsewhere in the genome[42], implying that EG1a also deserves ecospecies status.

Ecospecies evolution in the human stomach

The existence of these highly distinct genotypes within multiple geographical regions poses the question of how and when they have arisen and spread. Our explanation emphasizes continuity within their established primary host species, namely humans. The two ecospecies currently coexist stably within several human populations, and we propose that since the distinction between the types first arose in the ancestor of modern humans the types have coexisted continuously, as shown in Fig. 2a. This hypothesis can explain the current pattern of diversity parsimoniously, without requiring either host jumps into humans[3] or convergent evolution[43].

The ecospecies split is ancient. In the regions of the genome in which Hardy and Ubiquitous are distinct, the synonymous divergence between them is higher than that between any pair of strains from the same ecospecies (Extended Data Fig. 8), implying that the Hardy–Ubiquitous split is older than any known *H. pylori* population, including HpAfrica2, which originated in Khoisan groups in southern Africa and is thought to be the deepest-branching human population. Additional evidence for antiquity comes from dot plots showing many more genome rearrangements between pairs of strains in different ecospecies than between pairs within either of them (Extended Data Fig. 7), reflecting an ancient split in their clonal ancestry.

Despite the substantive differences between ecospecies, there is also good evidence of coexistence within the same human populations and parallel spread. First, in most of the genome, coexisting *H. pylori* from different ecospecies are more closely related to each other than they are to *H. pylori* from other geographical locations (Fig. 1b). Second, at the 100 genes that are differentiated between the ecospecies, those from Hardy and Ubiquitous strains have an approximately parallel history of spread, which incorporates Africa, Sahul, Siberia and North and South America, as can be seen from the similar topology and branch length of phylogenetic trees for these loci (Fig. 2b). A simple explanation for this approximately parallel history is that both types have been spread to these continents by the same, or similar, migrations of modern humans, before subsequent host jumps to big cats (*H. acinonychis*) and primates (UC Davis Primate Center isolates). These host jumps took place following the spread of *H. pylori* around the world and can be assigned to specific branches in Fig. 2a, but are probably tens of thousands of years old[4], because the resulting strains are quite distinct from any *H. pylori* in our sample. The geographical distribution of these lineages in the wild is unknown.

An alternative explanation for the origin of ecospecies is that they arose in an introgression event involving a different *Helicobacter* species or lineage. Our genomic data provide evidence against plausible introgression hypotheses. For example, if humans came out of Africa carrying Ubiquitous *H. pylori* and these hybridized with bacteria with Hardy alleles that had evolved in Neanderthals, Denisovans, big cats or aquatic mammals, this hybridization should have generated genetic exchange throughout the genome, as has been observed wherever distinct *H. pylori* populations coexist[26,33]. However, in the non-differentiated fraction of the genome we see no evidence for elongated branch length in those populations in which the Hardy ecospecies is found (Fig. 1b and Extended Data Fig. 2b) that would be generated by this gene flow, nor evidence of an elevated genetic distance to hpAfrica2 (Extended Data Fig. 8a). If the historical reservoir for Hardy was an archaic human or animal, it is also hard to explain the parallel history of spread shown in Fig. 2b, or indeed how these disparate locations were reached. Few species have been as successful in colonizing the world in the past 100,000 years as modern humans.

Although there is no evidence of import of diverged DNA into *H. pylori* populations, as required by an introgression hypothesis, there is good evidence of progressive divergence at the loci that define the ecospecies. According to the gene-by-gene topologies presented in Supplementary Fig. 1, most of the genes have separate Hardy and Ubiquitous clades. However, for 13 of 49 of these ecospecies-defining genes, hpAfrica2 strains cluster with *H. acinonychis*, implying that the divergence between Hardy and Ubiquitous versions commenced following the split from the hpAfrica2 population—in other words, within the history of extant human populations. Furthermore, there is evidence that positive selection has precipitated divergence, in the form of elevated dN/dS levels in comparisons between ecospecies, at ecospecies-defining genes (Extended Data Fig. 8d).

Notwithstanding the evidence for progressive evolution of Hardy and Ubiquitous ecospecies within humans, important uncertainties remain. The first is the age of the ecospecies split. Previous analysis[34] suggested that hpAfrica2 split from other *H. pylori* around 100,000 years ago and concluded that there is no direct evidence for *H. pylori* in humans before that. Assuming that divergence in dS values between hpAfrica2 and *H. pylori* Hardy strains at differentiated regions of the genome (Extended Data Fig. 8b,d, Hardy *H. pylori* subplot) accumulated according to a molecular clock implies that Hardy and Ubiquitous evolved at least 0.34/0.23 before this split, or at least 150,000 years ago. However, an alternative hypothesis is that hpAfrica2 strains codiverged with Khoisans, with a split time from other human populations of around 200,000 years—albeit with considerable uncertainty[35,44]. If this is used instead as a calibration point for Hardy evolution, this date estimate is pushed back by a factor of two to 300,000 years. Furthermore, residual gene flow between Hardy and Ubiquitous strains within these differentiated regions could result in substantial underestimation of when divergence between ecospecies commenced. Evidence for such gene flow is that dS values between Hardy and Ubiquitous strains are lowest between strains from the same geographic region (Extended Data Fig. 8d, Hardy *H. pylori* subplot).

Functional niche of Hardy ecospecies

A second major uncertainty relates to how Hardy strains have been able to reach Sahul, North and South America and Siberia from Africa. One speculative factor that could explain ecospecies distribution is variation in human diets during our colonization history. Hardy strains share an iron-cofactored urease with *Helicobacter* from carnivores. In mammals, the main source of nickel—the cofactor of urease in Ubiquitous

strains—is thought to be plant based[37]. This poses a challenge in regard to *Helicobacter* colonizing the stomachs of carnivores. Iron urease is thought to facilitate persistent colonization by species such as *H. felis* and *H. mustelae* even when nickel is limiting[38]. Further evidence for the potential role of diet is that several other genes that are differentiated between Hardy and Ubiquitous strains encode iron and nickel acquisition factors (*frpB4*, *tonB1*, *exbB*, *exbD*) or iron/acid adaptation regulators (*fur*, *arsS*).

Humans themselves have come under selection pressure due to changing diet. Hardy strains are currently found in human populations in which few crops are grown and that have ancestral alleles at the fatty acid desaturase locus[45], responsible for fatty acid metabolism, that have been strongly selected against within European populations over the past 10,000 years[46]. The modern distribution of *H. pylori* ecospecies could be explained if humans had relied principally on hunting when colonizing new locations but that this depleted large prey[47], leading to a dietary shift. In this scenario, the first human migrations could have transmitted both ecospecies but Hardy could then have died out in most locations along its migration path, long before the introduction of agriculture. If this scenario is correct, it implies that the ancestral population in which Hardy and Ubiquitous first differentiated was also heavily dependent on hunting. Functional work will be necessary to establish the conditions in which Hardy and Ubiquitous strains are able to coexist, and to further flesh out these hypotheses.

There is an intimate relationship between *H. pylori* pathogenesis and host iron status, and nutritional adaptation has also been proposed as one explanation for the 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[3], and *H. acinonychis* causes severe gastritis and is a frequent cause of cheetah death in captivity[48]. However, it remains to be seen whether the Hardy ecospecies can also be defined as a 'pathotype'. In addition to genes associated with metal uptake, a large fraction of the genes that are differentiated between ecospecies are in outer membrane proteins, including adhesins, with other genes being associated with virulence. This association is fascinating because it could shed light on the interactions among human diet, bacterial colonization strategies and virulence.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07991-z.

1. Suerbaum, S. & Michetti, P. *Helicobacter pylori* infection. *N. Engl. J. Med.* **347**, 1175–1186 (2002).
2. Amieva, M. & Peek, R. M. Jr. Pathobiology of *Helicobacter pylori*-induced gastric cancer. *Gastroenterology* **150**, 64–78 (2016).
3. Kersulyte, D. et al. Complete genome sequences of two *Helicobacter pylori* strains from a Canadian Arctic Aboriginal community. *Genome Announc*. https://doi.org/10.1128/genomeA.00209-15 (2015).
4. 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).
5. Ailloud, F. et al. Within-host evolution of *Helicobacter pylori* shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nat. Commun.* **10**, 2273 (2019).
6. Berthenet, E. et al. A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol.* **16**, 84 (2018).
7. Blanchard, T. G. et al. Genome sequences of 65 *Helicobacter pylori* strains isolated from asymptomatic individuals and patients with gastric cancer, peptic ulcer disease, or gastritis. *Pathog. Dis.* **68**, 39–43 (2013).
8. Camorlinga-Ponce, M. et al. Phenotypic and genotypic antibiotic resistance patterns in *Helicobacter pylori* strains from ethnically diverse population in Mexico. *Front. Cell. Infect. Microbiol.* **10**, 539115 (2020).
9. Chua, E. G. et al. Analysis of core protein clusters identifies candidate variable sites conferring metronidazole resistance in *Helicobacter pylori*. *Gastroenterol. Rep.* **7**, 42–49 (2019).
10. Didelot, X. et al. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc. Natl Acad. Sci. USA* **110**, 13880–13885 (2013).
11. Gutierrez-Escobar, A. J., Trujillo, E., Acevedo, O. & Bravo, M. M. Phylogenomics of Colombian *Helicobacter pylori* isolates. *Gut Pathog.* **9**, 52 (2017).
12. Hu, L. et al. Long-read- and short-read-based whole-genome sequencing reveals the antibiotic resistance pattern of *Helicobacter pylori*. *Microbiol. Spectr.* **11**, e0452222 (2023).
13. Kojima, K. K. et al. Population evolution of *Helicobacter pylori* through diversification in DNA methylation and interstrain sequence homogenization. *Mol. Biol. Evol.* **33**, 2848–2859 (2016).
14. Mehrotra, T. et al. Antimicrobial resistance and virulence in *Helicobacter pylori*: genomic insights. *Genomics* **113**, 3951–3966 (2021).
15. Moodley, Y. et al. *Helicobacter pylori*'s historical journey through Siberia and the Americas. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2015523118 (2021).
16. Munoz, A. B., Trespalacios-Rangel, A. A. & Vale, F. F. An American lineage of *Helicobacter pylori* prophages found in Colombia. *Helicobacter* **26**, e12779 (2021).
17. Munoz-Ramirez, Z. Y. et al. Whole genome sequence and phylogenetic analysis show *Helicobacter pylori* strains from Latin America have followed a unique evolution pathway. *Front. Cell. Infect. Microbiol.* **7**, 50 (2017).
18. Munoz-Ramirez, Z. Y. et al. A 500-year tale of co-evolution, adaptation, and virulence: *Helicobacter pylori* in the Americas. *ISME J.* **15**, 78–92 (2021).
19. Phuc, B. H. et al. *Helicobacter pylori* type 4 secretion systems as gastroduodenal disease markers. *Sci. Rep.* **11**, 4584 (2021).
20. Rehvathy, V. et al. Multiple genome sequences of *Helicobacter pylori* strains of diverse disease and antibiotic resistance backgrounds from Malaysia. *Genome Announc*. https://doi.org/10.1128/genomeA.00687-13 (2013).
21. Saranathan, R. et al. *Helicobacter pylori* infections in the Bronx, New York: surveying antibiotic susceptibility and strain lineage by whole-genome sequencing. *J. Clin. Microbiol.* https://doi.org/10.1128/JCM.01591-19 (2020).
22. Shetty, V. et al. High primary resistance to metronidazole and levofloxacin, and a moderate resistance to clarithromycin in *Helicobacter pylori* isolated from Karnataka patients. *Gut Pathog.* **11**, 21 (2019).
23. Thorell, K. et al. Identification of a Latin American-specific BabA adhesin variant through whole genome sequencing of *Helicobacter pylori* patient isolates from Nicaragua. *BMC Evol. Biol.* **16**, 53 (2016).
24. Thorell, K. et al. Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. *PLoS Genet.* **13**, e1006546 (2017).
25. Thorell, K. et al. The *Helicobacter pylori* Genome Project: insights into *H. pylori* population structure from analysis of a worldwide collection of complete genomes. *Nat. Commun.* **14**, 8184 (2023).
26. Thorpe, H. A. et al. Repeated out-of-Africa expansions of *Helicobacter pylori* driven by replacement of deleterious mutations. *Nat. Commun.* **13**, 6842 (2022).
27. Tuan, V. P. et al. A next-generation sequencing-based approach to identify genetic determinants of antibiotic resistance in Cambodian *Helicobacter pylori* clinical isolates. *J. Clin. Med.* https://doi.org/10.3390/jcm8060858 (2019).
28. Tuan, V. P. et al. Genome-wide association study of gastric cancer- and duodenal ulcer-derived *Helicobacter pylori* strains reveals discriminatory genetic variations and novel oncoprotein candidates. *Microb Genom*. https://doi.org/10.1099/mgen.0.000680 (2021).
29. Vale, F. F. et al. Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. *Sci. Rep.* **7**, 42471 (2017).
30. You, Y. et al. Genomic differentiation within East Asian *Helicobacter pylori*. *Microb. Genom*. https://doi.org/10.1099/mgen.0.000676 (2022).
31. Zhang, S. et al. Mutations of *Helicobacter pylori* RdxA are mainly related to the phylogenetic origin of the strain and not to metronidazole resistance. *J. Antimicrob. Chemother.* **75**, 3152–3155 (2020).
32. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
33. Falush, D. et al. Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585 (2003).
34. Moodley, Y. et al. Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.* **8**, e1002693 (2012).
35. Fan, S. et al. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923–939 (2023).
36. Kennemann, L. et al. *Helicobacter pylori* genome evolution during human infection. *Proc. Natl Acad. Sci. USA* **108**, 5033–5038 (2011).
37. Stoof, J. et al. Inverse nickel-responsive regulation of two urease enzymes in the gastric pathogen *Helicobacter mustelae*. *Environ. Microbiol.* **10**, 2586–2597 (2008).
38. 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. USA* **108**, 13095–13099 (2011).
39. Smet, A. et al. Macroevolution of gastric *Helicobacter* species unveils interspecies admixture and time of divergence. *ISME J.* **12**, 2518–2531 (2018).
40. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **361**, 1929–1940 (2006).
41. Cui, Y. et al. The landscape of coadaptation in *Vibrio parahaemolyticus*. *eLife* https://doi.org/10.7554/eLife.54136 (2020).
42. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. Preprint at *bioRxiv* https://doi.org/10.1101/385336 (2019).
43. Gressmann, H. et al. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genet.* **1**, e43 (2005).
44. Hollfelder, N., Breton, G., Sjodin, P. & Jakobsson, M. The deep population history in Africa. *Hum. Mol. Genet.* **30**, R2–R10 (2021).
45. Mathieson, I. Limited evidence for selection at the FADS locus in Native American populations. *Mol. Biol. Evol.* **37**, 2029–2033 (2020).
46. Mathieson, I. et al. Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. *Nat. Hum. Behav.* **7**, 790–801 (2023).
47. Dembitzer, J., Barkai, R., Ben-Dor, M. & Meiri, S. Levantine overkill: 1.5 million years of hunting down the body size distribution. *Quat. Sci. Rev.* **276**, 107316 (2022).

48. Kusters, J. G., van Vliet, A. H. & Kuipers, E. J. Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol. Rev.* **19**, 449–490 (2006).
49. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).

Helicobacter Genomics Consortium

Hafeza Aftab[13], Lotay Tshering[14], Dhakal Guru Prasad[14], Evariste Tshibangu-Kabamba[15], Ghislain Disashi Tumba[15], Patrick de Jesus Ngoma-Kisoko[16], Antoine Tshimpi-Wola[16], Dieudonné Mumba Ngoyi[16,17], Pascal Tshiamala Kashala[18], Modesto Cruz[19], José Jiménez Abreu[19], Celso Hosking[20], Jukka Ronkainen[21], Pertti Aro[22], Titong Sugihartono[3], Ari Fahrial Syam[23], Langgeng Agung Waskito[3], Hasan Maulahela[23], Yudith Annisa Ayu Rezkitha[24], Shaho Negahdar Panirani[7], Hamid Asadzadeh Aghdaei[25], Mohammad Reza Zali[26], Nasrin Mirzaei[7], Saeid Latifi-Navid[27], Takeshi Matsuhisa[28], Phawinee Subsomwong[2,29], Hideo Terao[30], Batsaikhan Saruuljavkhlan[2], Tadashi Shimoyama[31], Nagisa Kinjo[32], Fukunori Kinjo[33], Kazunari Murakami[34], Thein Myint[35], Than Than Aye[36], New Ni[37], Than Than Yee[38], Kyaw Htet[38], Pradeep Krishna Shrestha[39], Rabi Prakash Sharma[39], Jeewantha Rathnayake[40], Meegahalande Durage Lamawansa[40], Emilio Rudbeck[41], Lars Agreus[42], Anna Andreasson[43], Lars Engstrand[44], Varocha Mahachai[45], Thawee Ratanachu-Ek[46], Kammal Kumar Pawa[47], Tran Thi Huyen Trang[48], Tran Thanh Binh[49], Vu Van Khien[47], Ho Dang Quy Dung[5] & Dou Narith[50]

[13]Department of Gastroenterology, Dhaka Medical College and Hospital, Dhaka, Bangladesh. [14]Jigme Dorji Wangchuk National Referral Hospital, Thimphu, Bhutan. [15]University of Mbujimayi, Mbujimayi, Democratic Republic of the Congo. [16]University of Kinshasa, Kinshasa, Democratic Republic of the Congo. [17]National Institute of Biomedical Research (INRB), Kinshasa, Democratic Republic of the Congo. [18]Gastroenterology Service, Astryd Clinics, Kinshasa, Democratic Republic of the Congo. [19]Instituto de Microbiología y Parasitología (IMPA), Universidad Autónoma de Santo Domingo, Santo Domingo, Dominican Republic. [20]Universidad Autónoma de Santo Domingo, Santo Domingo, Dominican Republic. [21]Center for Life Course Health Research, University of Oulu, Finland and Primary Health Care Center, Tornio, Finland. [22]Arokero Oy, Tornio, Finland. [23]University of Indonesia, Jakarta, Indonesia. [24]Faculty of Medicine, Muhammadiyah University of Surabaya, Surabaya, Indonesia. [25]Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [26]Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [27]Department of Biology, Faculty of Sciences, University of Mohaghegh Ardabili, Ardabil, Iran. [28]Nippon Medical School, Tokyo, Japan. [29]Department of Microbiology and Immunology, Hirosaki University Graduate School of Medicine, Hirosaki, Japan. [30]Oita University, Oita, Japan. [31]Aomori General Health Examination Center, Aomori, Japan. [32]Ryusei Hospital, Naha, Japan. [33]Center for Gastroenterology, Urasoe General Hospital, Urasoe, Japan. [34]Department of Gastroenterology, Oita University Faculty of Medicine, Yufu, Japan. [35]Department of Gastroenterology, Yangon General Hospital, University of Medicine, Yangon, Myanmar. [36]Department of Gastroenterology, Thingangyun Sanpya General Hospital, University of Medicine (2), Thingangyun, Myanmar. [37]Department of Gastroenterology, Mandalay General Hospital, University of Medicine, Mandalay, Myanmar. [38]No. 1 Defense Service General Hospital (1000 Bedded), Yangon, Myanmar. [39]Gastroenterology Department, Maharajgunj Medical Campus, Tribhuvan University Hospital, Kathmandu, Nepal. [40]Department of Surgery, Teaching Hospital Peradeniya, University of Peradeniya, Kandy, Sri Lanka. [41]Clinical Genomics Gothenburg, Bioinformatics and Data Centre, University of Gothenburg, Gothenburg, Sweden. [42]Division of Family Medicine, Karolinska Institutet, Stockholm, Sweden. [43]Stress Research Institute, Department of Psychology, Faculty of Social Sciences, Stockholm University, Stockholm, Sweden. [44]Center for Translational Microbiome Research, Department for Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. [45]GI and Liver Center, Bangkok Medical Center, Bangkok, Thailand. [46]Department of Surgery, Rajavithi Hospital, Bangkok, Thailand. [47]Department of Medicine, Chulabhorn International College of Medicine (CICM) at Thammasat University, Pathumthani, Thailand. [48]108 Military Central Hospital, Hanoi, Vietnam. [49]Tam Anh General Hospital, Ho Chi Minh City, Vietnam. [50]Department of Endoscopy, Cho Ray Phnom Penh Hospital, Phnom Penh, Cambodia.

Methods

Genome collection

We collected a total of 9,188 *Helicobacter* whole-genome sequences from public and private sources, including 4,210 *H. pylori* and one *H. acinonychis* genomes publicly available in Enterobase[50] (as of 6 June 2022), 1,011 samples from the *Helicobacter pylori* genome project[25] (https://doi.org/10.5281/zenodo.10048320)[51] and 350 samples available in NCBI, BIGs and Figshare; and 3,615 *H. pylori* novel genomes from different geographic regions around the world and one novel *H. acinonychis*. The novel sequences included 2,133 isolates collected by Y.Y. at the Department of Environmental and Preventive Medicine, Faculty of Medicine, Oita University, Japan, 244 strains from Iran collected by A.Y. and 142 genomes from different parts of the world, including 89 from the Swedish Kalixanda cohort[52]. Lastly, 1,096 worldwide DNA samples and one *H. acinonychis* sample were contributed by Mark Achtman. These sequences have also been deposited in Enterobase at the following workspace: https://enterobase.warwick.ac.uk/a/108555.

Samples from the Yamaoka laboratory were sequenced at Novogene Co., Ltd., Beijing, China with the Illumina NOVA PE150 platform and assembled using the SPAdes genome assembler v.3.15.3 (ref. 53) by downsampling read depth to 100 base pairs, specifying a genome size of 1.6 megabase pairs and enabling the option --careful. Of those samples obtained from M.A., 916 were sequenced using either the Illumina MiSeq platform at the University of Gothenburg, Sweden or the Illumina NOVA PE150 platform at Novogene Co., Ltd, UK; a further 180 were sequenced at the University of Warwick, UK. Remaining sequences were sequenced on the Illumina MiSeq platform at Karolinska Institutet and the University of Gothenburg and assembled using the BACTpipe pipeline (https://doi.org/10.5281/zenodo.4742358)[54].

We then filtered out redundant genomes, defined as those sequences with SNP distance below 200, and removed low-quality genomes based on assembly fragmentation (over 500 contigs), coverage to the 26695 *H. pylori* reference strain (below 70%) and contamination (above 90% *H. pylori*) as predicted by Kraken v.2.1.2 (ref. 55), to obtain a final total of 6,866 genomes corresponding to the 2,916 genomes first reported in this work and 3,950 that are publicly available (Supplementary Table 2).

Sequence alignment and core genome variant calling

Single-nucleotide polymorphisms from the core genome (core SNPs) were called using an algorithm based on MUMmer v.3.20 (ref. 56), as previously described[57]. We first aligned each genome sequence of the entire dataset to the *H. pylori* 26695 reference strain (NC_000915.1) using nucmer v.3.1. Next, snp-sites v.2.5.2 was used to call all variants from the whole-genome alignment obtained. Variants present in at least 99% of genomes were finally extracted using VCFtools v.0.1.17, generating a total of 866,840 core SNPs.

Population assignment

To assign a population to the final dataset, we first defined a reference subset of 285 strains consistently assigned into one of 19 *H. pylori* populations/subpopulations in previous reports[15,33,58,59] and that we confirmed by running fineSTRUCTURE[32] 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. This subset was then considered as a donor panel to paint each sample of the entire dataset using ChromoPainter v.2 (ref. 32). Genomes were labelled with a population based on the largest fraction of their genome painted by a population.

PCA

PCA on the whole dataset was performed using SNPs extracted from the global alignment file, following linkage disequilibrium pruning to remove linked SNPs (window size, 50 base pairs; step size, ten variants; r^2 threshold, 0.1), using the software PLINK (v.1.9)[60].

The analysis and plotting (for this section and the following) were done using R v.4.3.1 and python v.3.10.6, as well as the R packages ggplot2 v.3.3.6 and tidyverse v.1.3.2 and python library numpy v.1.23.2.

Phylogenetic analysis

For reconstruction of the various phylogenetic trees, we used coding sequences that were aligned with strain 26695 (see above). When looking at specific genes (*vacA*, *ureA*, *ureB*), gene sequences were first obtained from the individual strains annotation file then aligned using MAFFT (v.7.505, option --auto)[61]. The tree shown in Extended Data Fig. 2a was built using SNPs from all coding sequences. The trees shown in Fig. 1b,c and Extended Data Fig. 2b,c were built using SNPs from the undifferentiated (B panels) and differentiated (C panels) genes, respectively. The trees shown in Supplementary Figs. 1–4 were built using SNPs from specific genes. In addition, for *ureA* (Supplementary Fig. 3) and *ureB* (Supplementary Fig. 4), the sequences were separated into two types because some strains had two copies of the gene. The choice of copy type was based on the similarity between sequences (based on tree clustering and BLAST results, in particular against *H. cetorum*). Using the various alignments of nucleotide sequences, maximum-likelihood trees were constructed using FastTree software[62] (v.2.1.10, option -nt). The trees were then rooted based on a given outgroup normally used for *H. pylori*: hpAfrica2 and *H. acinonychis* for trees looking at the whole genome or at undifferentiated genes, and Hardy strains for trees looking at differentiated genes. In the case of those trees looking at individual genes, we used *H. cetorum* as an outgroup, these genes having been chosen after their sequences had been BLASTED against the *H. cetorum* genome (see below). Rooting was done using the R package ape[63] (v.5.7-1, root function). Plotting was done using the R package ggtree v.3.2.1.

The population-level trees shown in Fig. 2b were built via a neighbour-joining algorithm (R package ape, function nj), using a matrix of the average distance between populations represented in both Hardy and Ubiquitous ecospecies. As an equivalent of *H. acinonychis* (Hardy), we used strains from hpAfrica2. Distances between strains were calculated using the dist.dna function (option model, 'raw') from the ape R[64] package. The trees were rooted with hpAfrica2/*H. acinonychis* as outgroups, using the same root function as previously.

FineSTRUCTURE

To further investigate the population structure of Hardy and Ubiquitous strains, we analysed 295 strains assigned by ChromoPainter to hspIndigenousSAmerica, hspSiberia and hspIndigenousNAmerica *H. pylori* populations by running fineSTRUCTURE with 200,000 iterations of burn-in and Markov chain Monte Carlo, using as input the haplotype data prepared with SNPs from the core genome of those 295 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 244 strains from hspSiberia and hspIndigenousNAmerica, we performed GWAS to determine which biallelic core SNPs were significantly associated with the ecospecies. Although Hardy strains were also found in hspIndigenousSAmerica, we chose to remove this population from GWAS analysis due to the small number (two of 49) of Hardy strains. Ubiquitous strains were coded as 0 (198 strains) and Hardy strains as 1 (46 strains). GWAS was performed using the R package bugwas (v.0.0.0.9000)[65], which considers the population structure using PCA, followed by GEMMA (v.0.93) to perform GWAS analysis. Using a standard significance threshold of $-\log(P) = 5$, 4,609 out of 285,792 core biallelic SNPs were significantly associated with the Hardy clade.

To reinforce the results obtained by GWAS, we calculated per-site FST between Ubiquitous and Hardy ecospecies with the R package PopGenome[66] using the same set of strains and SNPs. We considered

a SNP to be differentiated between Hardy and Ubiquitous ecospecies if it was significantly associated with the ecospecies by GWAS ($-\log(P) > 10$), and highly differentiated between the two groups based on its FST value (FST > 0.9). Of the core biallelic SNPs we found 2,568 differentiated coding SNPs and 175 differentiated intergenic SNPs. We considered a SNP to be undifferentiated if $-\log(P) < 10$ and FS < 0.5 (265,621 coding and 8,950 intergenic SNPs). All other SNPs (7,756 coding and 591 intergenic) were considered intermediate. Following separation of SNPs into three classes, we also distinguished three types of gene based on those present in the 26695 genome: differentiated (100 genes; Supplementary Table 3), containing at least five differentiated SNPs; undifferentiated (1,034 genes), with only undifferentiated SNPs; and the remaining genes (443 genes), which we considered 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 with both the genetic distance to Ubiquitous strains and the number of Hardy blocks. For a given strain, the distance to Ubiquitous strains is the average number of differences between the sequences of this strain and sequences of Ubiquitous strains from hspIndigenousSAmerica, hspSiberia and hspIndigenousNAmerica. The sequences were those aligned on the 26695 sequence, and gaps were removed.

Hardy blocks were defined based on differentiated SNPs: for each strain, if two adjacent differentiated SNPs had the same allele and were part of the same gene, we considered that they were part of the same Hardy block; otherwise, they were from different blocks.

Pangenome analysis

First, we estimated the gene content of each sample with prokka v.1.4.6 software[67] using the proteome of 26695 as reference. Then, .gff files were used as input for Panaroo's v.1.2.8 (ref. 68) pangenome pipeline using the strict mode, merging paralogues based on sequence identity of 0.95, length difference of 0.90 and core threshold of 0.95. For this analysis, a smaller dataset was used that consisted of all strains from hspSiberia and hspIndigenousNAmerica (that is, all Hardy strains were included and all Ubiquitous strains from the same populations), as well as randomly chosen strains from the other populations (size of the sample dataset, 721 strains). Then, a hierarchical clustering based on the presence/absence of pangenes was conducted with the pheatmap v.1.0.12 package in R v.4.3.1 using the complete linkage method.

For detection of *cagA*, *vacA* and *ureAB* homologues in diverse *Helicobacter* spp., non-*pylori Helicobacter* genomes were recovered from either GenBank or Enterobase (Supplementary Table 4) and annotated using Prokka. Strains/species were considered if they encoded an intact copy of UreA1B1 (indicating gastric tropism) and available metadata indicated isolation from humans or animals. Metagenome-derived genomes from non-animal hosts were excluded. Host diets were identified using Wikipedia.

For identification of putative homologues, pangenome analysis was performed together with a subset of *H. pylori* strains (Supplementary Table 2) using panaroo with 70% sequence identity and 75% sequence coverage cut-off. For *vacA*, this was supplemented with additional manual inspection (for example, incomplete genomes) using Mauve[69] and Tablet[70] and data from the literature (for example, *H. cetorum*[71]).

Comparison with *H. cetorum* and phylogenetic analysis of differentiated genes

We used *H. cetorum* as an outgroup in the study of differentiated genes, specifically *H. cetorum* strain MIT99-5656 (downloaded 14 February 2023 from NCBI (https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000259275.1/). Gene sequences were obtained from the GenBank file provided.

A phylogenetic analysis was performed on differentiated genes, with the *H. cetorum* genome included. First, we obtained *H. cetorum* gene sequences using BLASTing (blastn v.2.11.0)[72] of a Hardy and a Ubiquitous version of each differentiated gene against the *H. cetorum* genome. For those genes that returned at least one hit, the phylogenetic tree of the gene was generated using FastTree and rooted on the *H. cetorum* sequence (see above).

Comparison of genome structure

Significant structural variations, including inversions, gaps, repeats and gene cluster rearrangement, are readily visualized using a dot plot. For investigation of genome structure similarity and the difference between Hardy and Ubiquitous groups, we used the Gepard program[73] (v.1.40 jar file from https://github.com/univieCUBE/gepard) to make the dot plot. Different comparisons were considered: Hardy versus Hardy, Hardy versus Ubiquitous and Ubiquitous versus 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 colour 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 functional enrichment analysis using the website DAVID[74]. Three hypothetical proteins were removed because they lacked a unique identifier. The background set of genes chosen (see Supplementary Table 3 for a list of genes tested and their categories) for the comparison was based on those genes present in *H. pylori* 26695. A term was considered as significant at $P < 0.05$ following Benjamini correction for multiple tests.

Calculation of dN/dS and ANI

For each strain in the dataset, we estimated its dN/dS and dS values to an outgroup population using the Yang and Nielsen method (YN00) in PAML (v.4.9) software[75,76], whereas ANI was calculated with FastANI (v.1.34 (ref. 77)). The dN/dS, dS and ANI values shown in plots were averaged over pairwise comparisons with each of the different outgroup strains. Codons that coded for a stop in at least one of the strains were removed from all strains, and dN/dS and ANI values were calculated pairwise using the coding sequences from those aligned to the reference genome (26695, global alignment). In addition, we separated the values between undifferentiated and differentiated coding sequences. We used three different outgroups: hpAfrica2, *H. acinonychis* (Hardy strains) and *H. pylori* Hardy strains, with outgroup population/ecospecies not shown in the plots (only population/ecospecies of the 'focal' strain).

Ethics and inclusion statement

The *Helicobacter* genomics consortium includes gastroenterologists and researchers from several developing countries. Its aim is to characterize the genetic diversity of *Helicobacter* in human populations across the world, and its correlations with gastric disease. In Bangladesh, Bhutan, Congo DR, Dominican Republic, Indonesia, Japan, Myanmar, Nepal, Sri Lanka, Thailand and Vietnam, all preparation for endoscopy surveying was performed by local researchers (persons in the consortium and PhD students at Oita University). Ethical permission for the collection of human gastric biopsy material was obtained for all cohorts, including informed consent from participating individuals. Endoscopy was performed by local physicians and Y.Y. Culturing of bacteria, DNA extraction, next-generation sequencing and basic genetic analysis in these countries were performed at Oita University, principally by doctoral students who were locally recruited in several of the countries and are coauthors of the paper. In addition, the doctors

and scientists involved in this consortium are actively involved in the research process and are kept up to date with its findings. This training and dissemination programme will help to spread both genomics knowledge and best practice for treating gastric illness, from Japan, where there has been considerable success in mitigating the burden of gastric cancer and other conditions, to other less developed nations. The study protocol (Iranian strains) was approved by the Institutional Ethical Review Committee of the Research Institute for Gastroenterology and Liver Diseases at Shahid Beheshti University of Medical Sciences, Tehran, Iran (no. IR.SBMU.RIGLD.REC.1395.878). All experiments were performed in accordance with relevant guidelines and regulations recommended by the institution. Written informed consent was obtained from all enroled subjects and/or their legal guardians before sample collection. The sampling of DNA used to generate the new Siberian genomes has been detailed previously[15]. The remaining new genomes were also sourced from previously collected *H. pylori* DNA (for example, refs. 33,34,59). The study protocol for the Swedish Kalixanda genomes was approved by Umeå University ethics committee, and the study was conducted in accordance with the Helsinki Declaration. The science of microbiology has not generally viewed human-derived microbes as belonging to the individuals they came from and routine publication of genetic sequences from bacterial pathogens has enabled many public health applications. In a few bacteria, however, *Helicobacter pylori* being one of them, the tight coupling between human and bacterial population structure makes unexpected inferences about human hosts possible from bacterial data, which may be far from the uses envisioned when consent was obtained for sample collection. We will strive to ensure that the design of future studies is built around the needs of communities in which they are performed and that consent procedures provide accurate information on how the samples will be used, informed by recent scientific advances. Our finding of a highly distinct Hardy ecospecies has potential implications for infected individuals in many Indigenous communities known to have a high gastric disease burden. However, the pathogenicity profile, either in single or mixed infection, is currently unknown. In keeping with the TRUST code of conduct, we are maintaining and developing contacts with researchers working in communities where Hardy strains have been isolated with the intention of consulting representatives of these communities to ascertain and accommodate their interests in collaborative research on the functional characterization of these strains, including associations with gastric disease.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The list of strains used is provided in Supplementary Tables 1 and 2, with NCBI accession numbers for all newly sequenced strains given in Supplementary Table 1. The full dataset is also available on the Enterobase worksheet (https://enterobase.warwick.ac.uk/a/108555).

Code availability

The scripts for the following analyses—PCA, phylogenetic analysis, GWAS, FST and dN/dS—are available at GitHub (https://github.com/EliseTourrette/HpEcospecies_Tourrette2023) and at Zenodo (https://doi.org/10.5281/zenodo.12740447)[78].

50. 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).
51. Wang, D. et al. The *Helicobacter pylori* Genome Project (HpGP) Phase1 dataset and 255 *H. pylori* population reference dataset. Zenodo https://doi.org/10.5281/zenodo.10048320 (2023).
52. Aro, P. 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).
53. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
54. Rudbeck, E. et al. ctmrbio/BACTpipe: BACTpipe v3.1.0. Zenodo https://doi.org/10.5281/zenodo.4742358 (2021).
55. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
56. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
57. Cui, Y. et al. Epidemic clones, oceanic gene pools, and Eco-LD in the free living marine pathogen *Vibrio parahaemolyticus*. *Mol. Biol. Evol.* **32**, 1396–1410 (2015).
58. Moodley, Y. & Linz, B. *Helicobacter pylori* sequences reflect past human migrations. *Genome Dyn.* **6**, 62–74 (2009).
59. Linz, B. et al. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918 (2007).
60. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
61. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
62. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
63. Paradis, E. & Schliep, K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
64. R Core Team. *R: a Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
65. Earle, S. G. et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).
66. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
68. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
69. Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
70. Milne, I. et al. Tablet–next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
71. 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).
72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
73. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
74. 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).
75. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
76. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
77. Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
78. Elise. EliseTourrette/HpEcospecies_Tourrette2023: Publication. Zenodo https://doi.org/10.5281/zenodo.12740447 (2024).

Acknowledgements We are grateful to the many research participants and health care professionals who facilitated sample collection efforts. We thank M. Achtman for providing strain DNA and J. Solnick for information on primate strains. This work is supported by the National Natural Science Foundation of China (nos. 32170640 and 32211550014) to D.F., 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 of Japan (nos. 18KK0266, 19H03473, 21H00346 and 22H02871 to Y.Y. and 21K08010 to T.M.), and by the Grants-in-Aid of the National Fund for Innovation and Development of Science and Technology from the Ministry of Higher Education Science and Technology of the Dominican Republic (nos. 2012-2013-2A1-65 and 2015-3A1-182 to M.C.). This work was also supported by the Japan Agency for Medical Research and Development and Japan Society for the Promotion of Science (Bilateral Programs, Japan–China, to Y.Y.), 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 (R.-K.V.). K.T. was supported by the Swedish Society for Medical Research, Assar Gabrielsson Foundation (nos. FB20-12 and FB21-89) and the Magnus Bergvall Foundation. Sequencing and database development was also funded in part by Wellcome Trust grant no. 202792/Z/16/Z to M. Achtman The computations and data storage required for assembly and annotation of genomes sequenced at the University of Gothenburg were enabled by resources in project nos. snic-2021/22-229 and snic-2021/23-234 provided by the National Academic Infrastructure for Supercomputing in Sweden and the Swedish National Infrastructure for Computing at UPPMAX HPC, partially funded by the Swedish Research Council through grant agreement nos. 2022-06725 and 2018-05973. This project has been funded in part by federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services under contract no. 75N91019D00024 (D.W.). A.Y. was supported financially by research grant nos. RIGLD 722, 878, 968, 969 and 1128

from the Foodborne and Waterborne Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Author contributions Project supervision was undertaken by Y.Y., K.T. and D.F. Strain isolation, sequencing and genomics were the responsibility of R.C.T., T.M., M.M., K.A.F., R.I.A., R.-K.V., V.P.T., the *Helicobacter* Genomics Consortium, A.Y., L.M.O., Z.Z., Y.Y. and K.T. Evolutionary data analysis was carried out by E.T., R.C.T., S.L.S., K.T., D.W. and D.F. E.T., R.C.T., S.L.S., L.M.O., K.T. and D.F. wrote the manusript.

Funding Open access funding provided by University of Gothenburg.

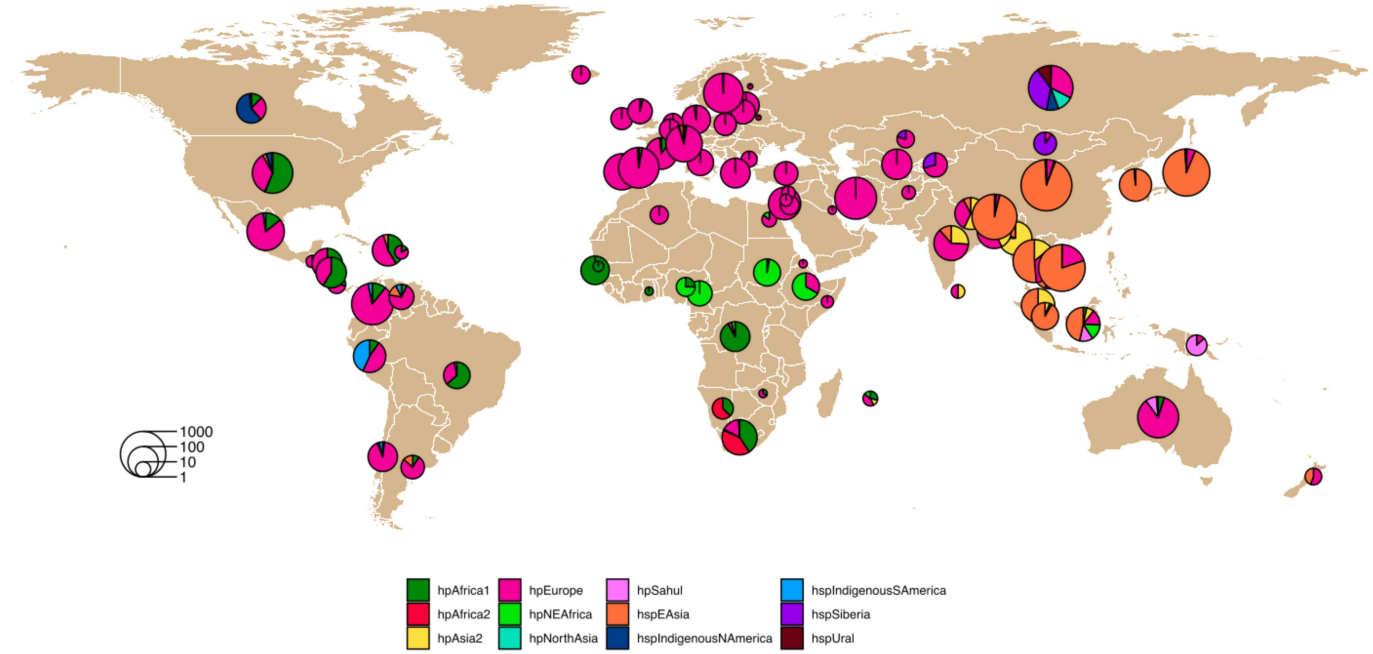
Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07991-z.

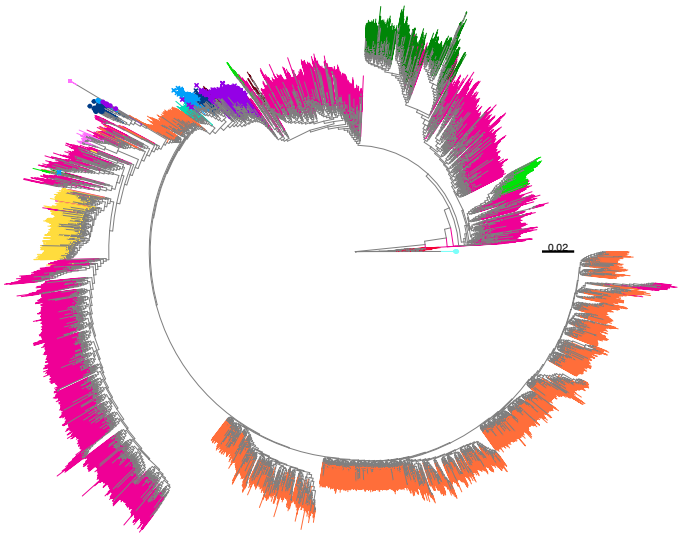
Correspondence and requests for materials should be addressed to Yoshio Yamaoka, Kaisa Thorell or Daniel Falush.

Peer review information *Nature* thanks Sébastien Calvignac-Spencer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Origin of the different strains of our global dataset. The size of the pie charts shows the number of isolates from the country, with areas scaling logarithmically with sample size. The pie charts show the proportion of isolates assigned to each *H. pylori* population.

a) Whole genome tree, rooted by *Helicobacter acinonychis*



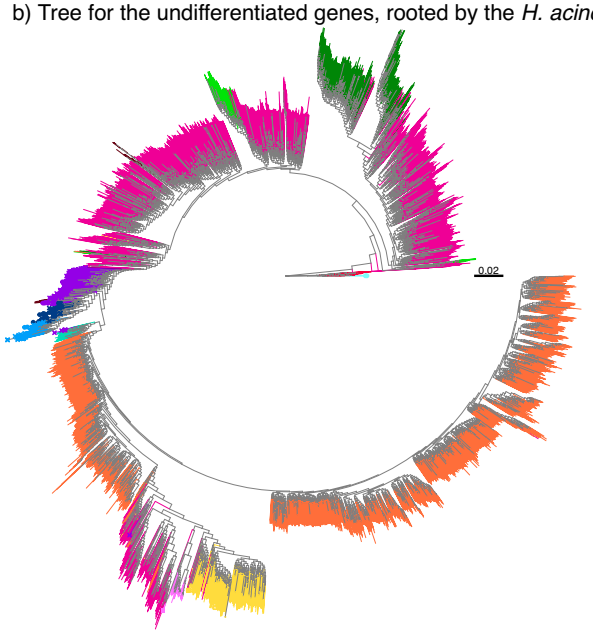
Population

- *H. acinonychis*
- hpAfrica1
- hpAfrica2
- hpAsia2
- hpEurope
- hpNEAfrica
- hpNorthAsia
- hpSahul
- hspEAsia
- hspIndigenousNAmerica
- hspIndigenousSAmerica
- hspSiberia
- hspUral

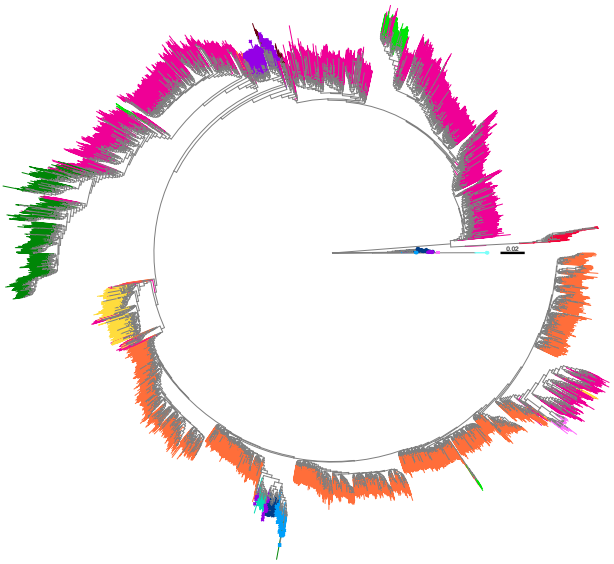
Ecotype

- Hardy (*H. acinonychis*)
- Hardy
- Ubiquitous (hspSiberia, hspIndigenousNAmerica, hspIndigenousSAmerica)
- Hardy (Primate)
- Ubiquitous (Primate)

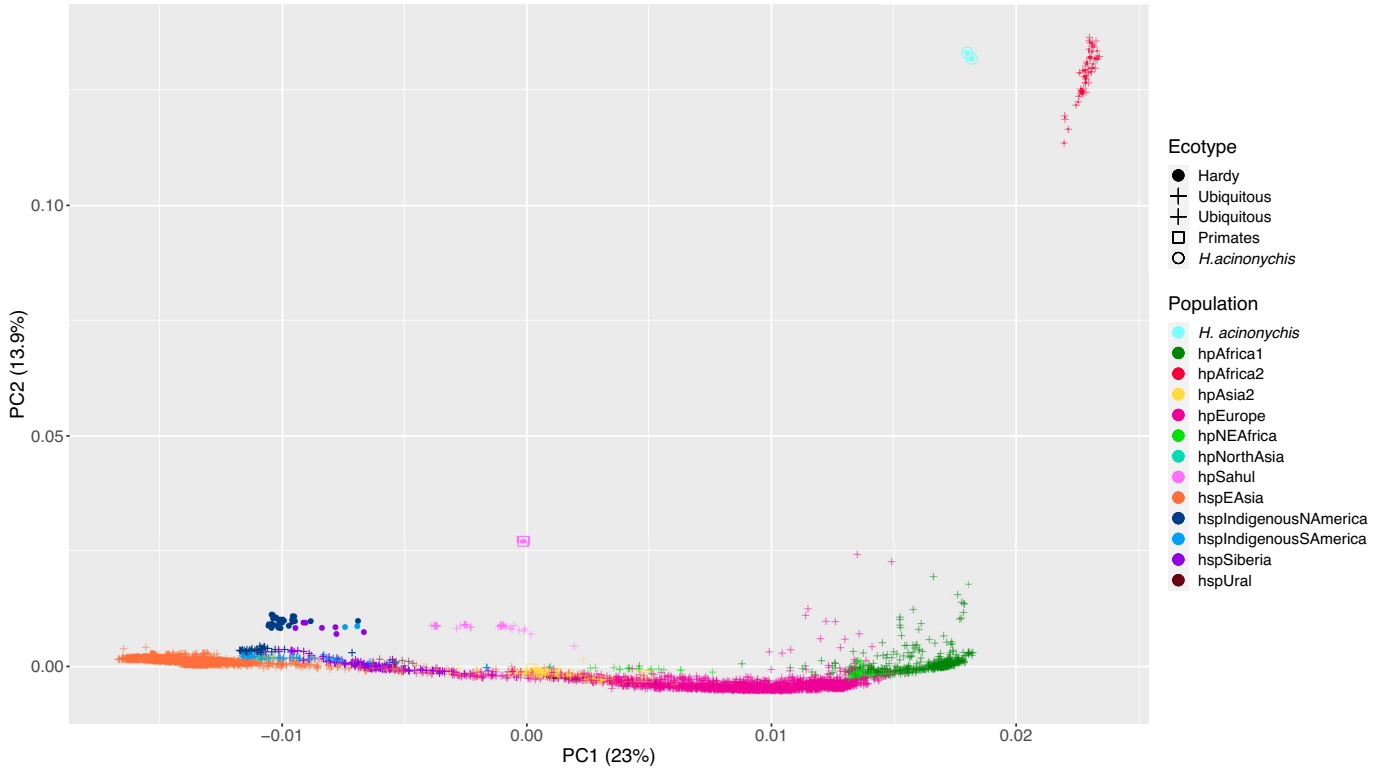
b) Tree for the undifferentiated genes, rooted by the *H. acinonychis*



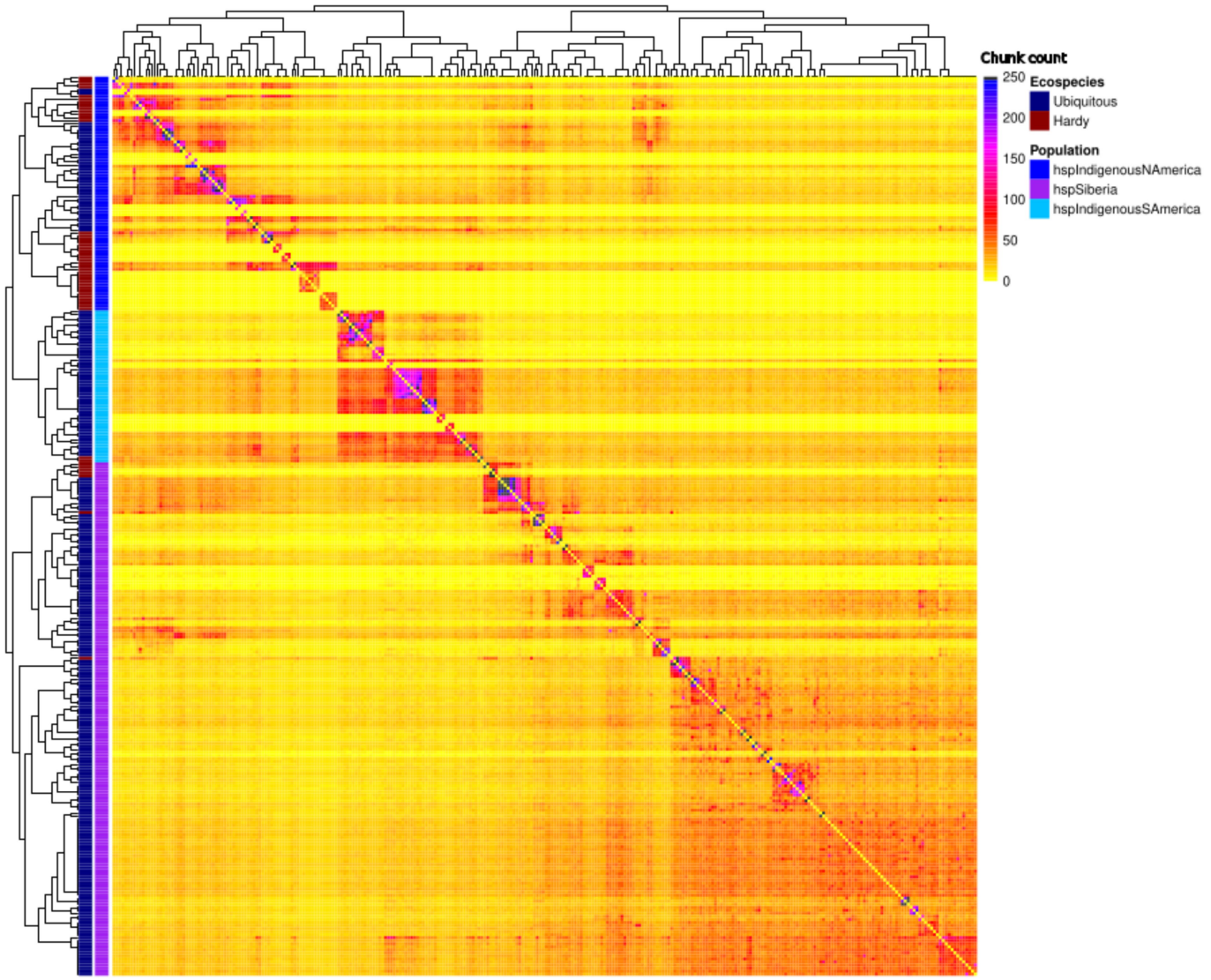
c) Tree for the differentiated genes, rooted by the Hardy strains



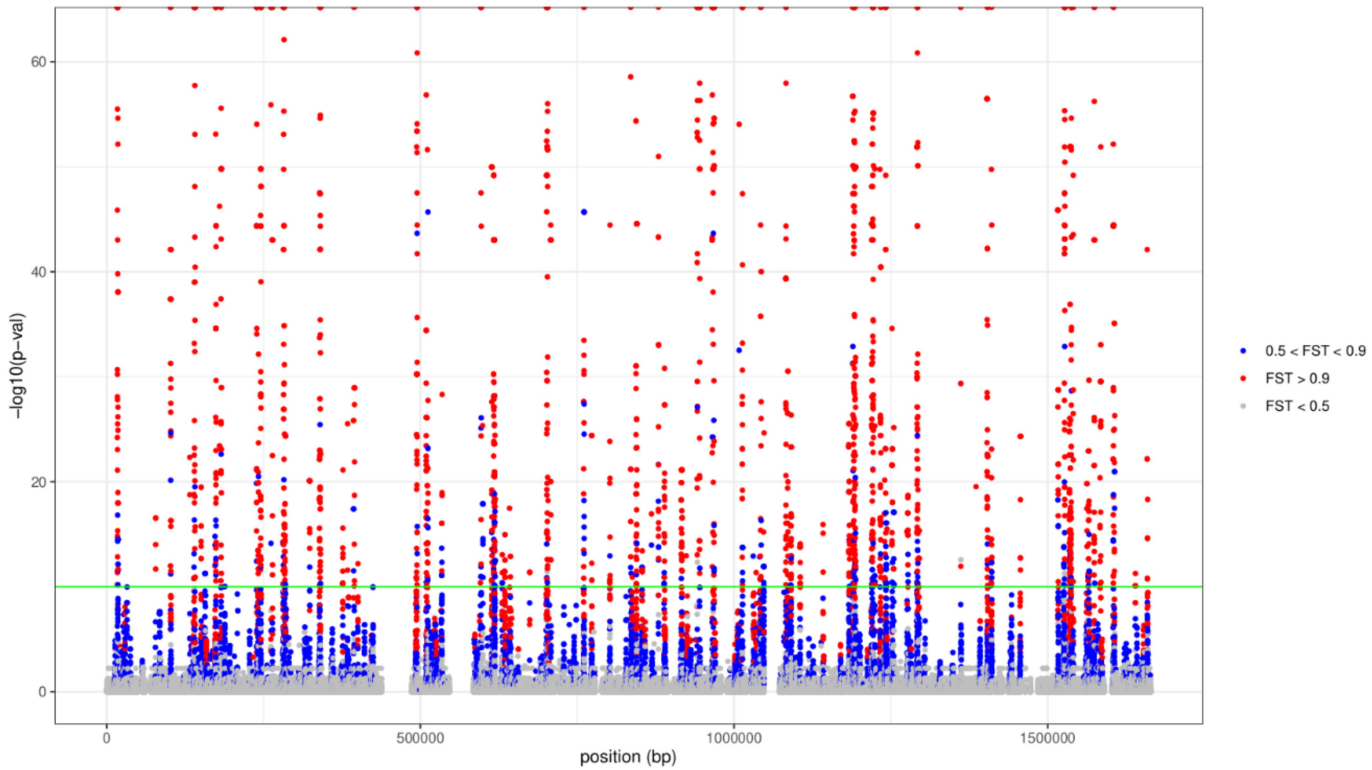
Extended Data Fig. 2 | Phylogenetic trees for all strains in the dataset. The branches are coloured based on their population. Hardy strains are represented with a circle while the other strains from the same populations are represented with a cross. The primate strains are represented with squares. Phylogenetic trees for all the genes (A). Some strains from hspIndigenousNAmerica, hspIndigenousSAmerica and hspSiberia do not cluster with their expected population. For undifferentiated genes (B) and for the differentiated genes (C). The branches are coloured based on the population and the strains from the Hardy clade are indicated with a dot.



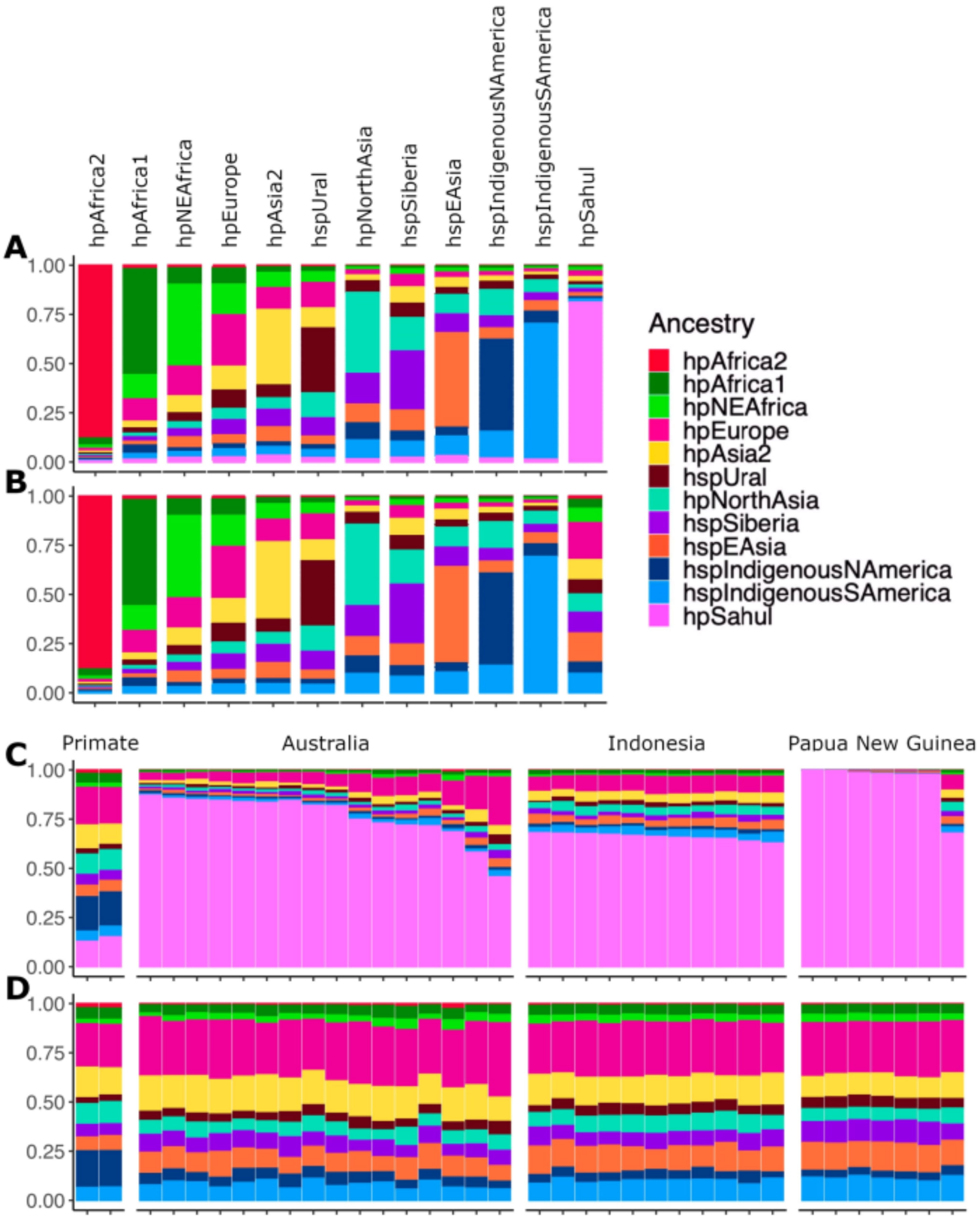
Extended Data Fig. 3 | First two components of the Principal Components Analysis (PCA) from the entire dataset. Strains are coloured based on their population and the strains from the Hardy ecospecies are represented by a dot while the crosses represent Ubiquitous strains. Squares and circles respectively indicate primate and *H. acinonychis* strains.



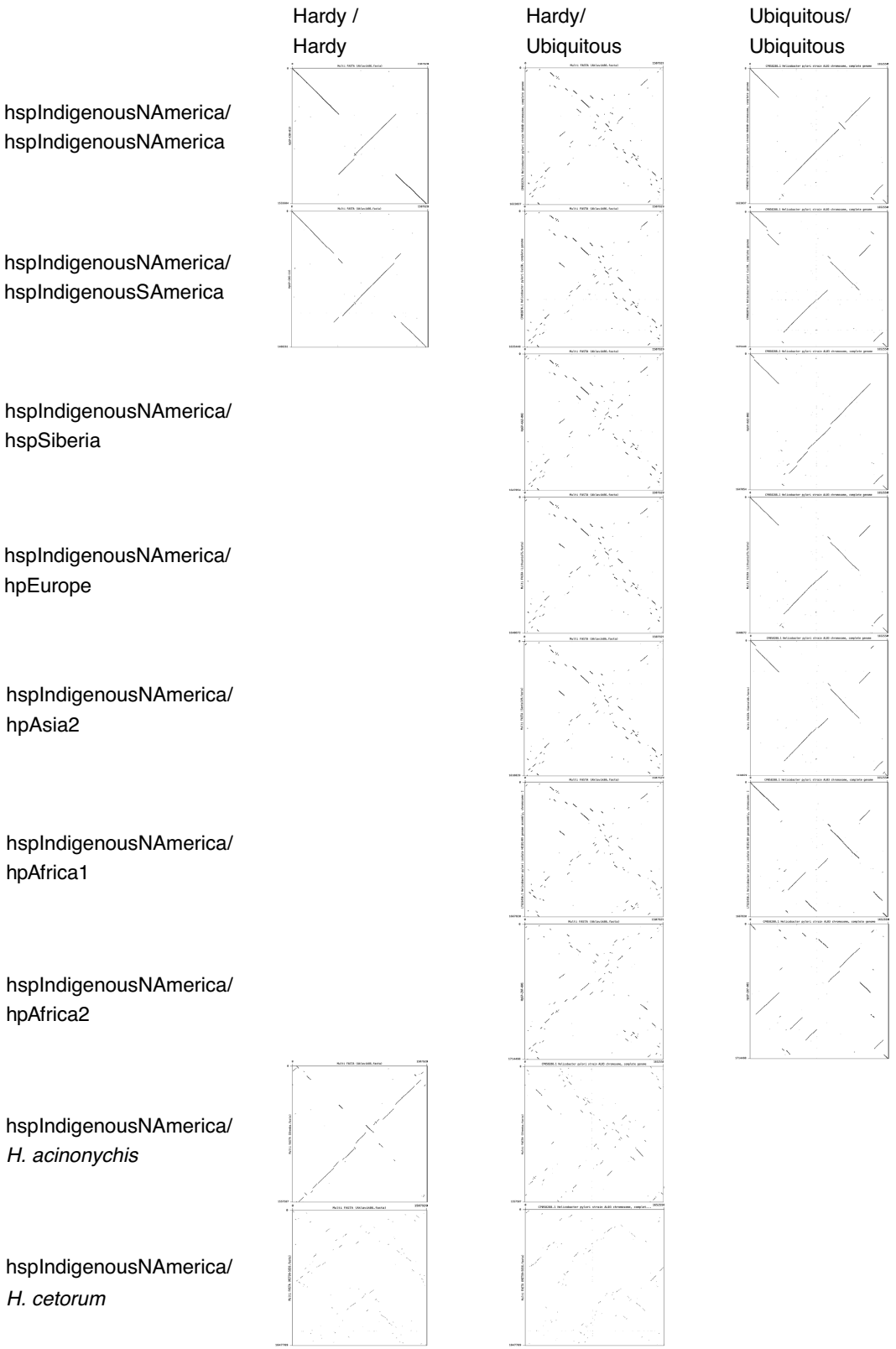
Extended Data Fig. 4 | FineSTRUCTURE analysis of the strains from hspSiberia, hspIndigenousNAmerica and hspIndigenousSAmerica. The strains from the Hardy clade are highlighted by red shading, overlaying the dendrogram on the left of the plot, while the Ubiquitous strains are highlighted with blue. FineSTRUCTURE uses an in silico chromosome painting algorithm to fit each strain as a mosaic of nearest neighbours, chosen from the other strains in the dataset. Each row shows the coancestry vector for one strain, which is a count of the number of segments of DNA used in the painting from each of the other strains in the dataset. High coancestry between strains implies that they are nearest neighbours for many segments of the genome and hence share genetic material from a common gene pool. FineSTRUCTURE based clustering is more sensitive to recent gene flow than clustering using genetic distances.



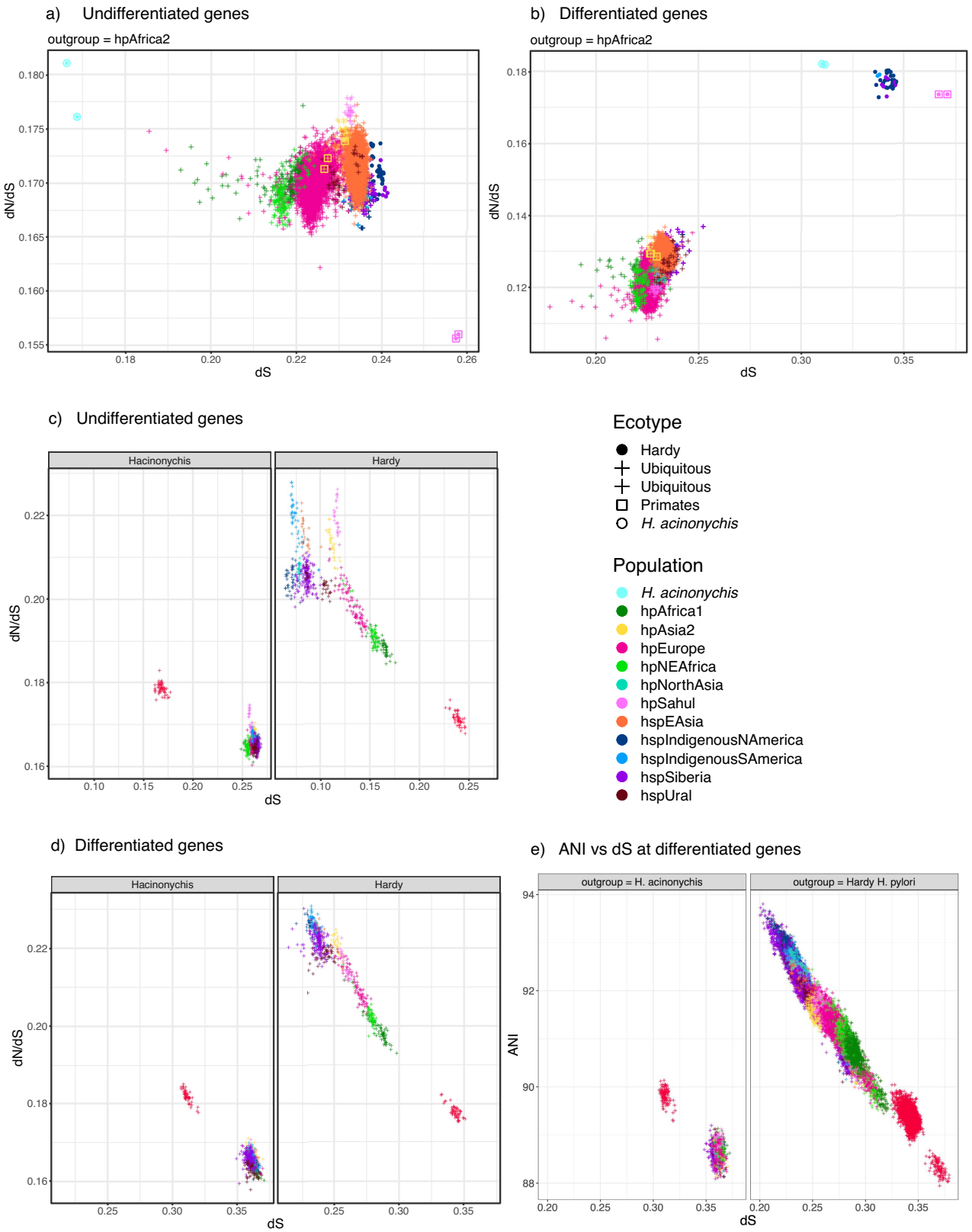
Extended Data Fig. 5 | Manhattan plot resulting from a GWAS analysis of the Hardy vs Ubiquitous strains from hspSiberia and hspIndigenousNAmerica. The green line represents the significance threshold (-log10(p) = 10, bayesian Wald test with a correction for multi-testing giving a significance threshold equals to $\alpha/n_{snps} = 0.05/285{,}792$). Points are coloured based on FST (fixation index between Hardy and Ubiquitous ecospecies) values (red: FST > 0.9, blue: FST 0.5 – 0.9, grey: FST < 0.5). Half points at the top of the plot indicate an estimated p-value of zero and FST of one.



Extended Data Fig. 6 | Average ancestry profiles of Global *H. pylori*. (A) With hpSahul donors and (B) without hpSahul donors. Close-up of Hardy primates and hpSahul strains (C) with hpSahul donors and (D) without hpSahul donors. Although the primate strains do not correspond to any hpSahul strains present in the data (based on their ancestry profile in the presence of an hpSahul donor), they can still be assigned to hpSahul (based on their ancestry profile without an hpSahul donor).

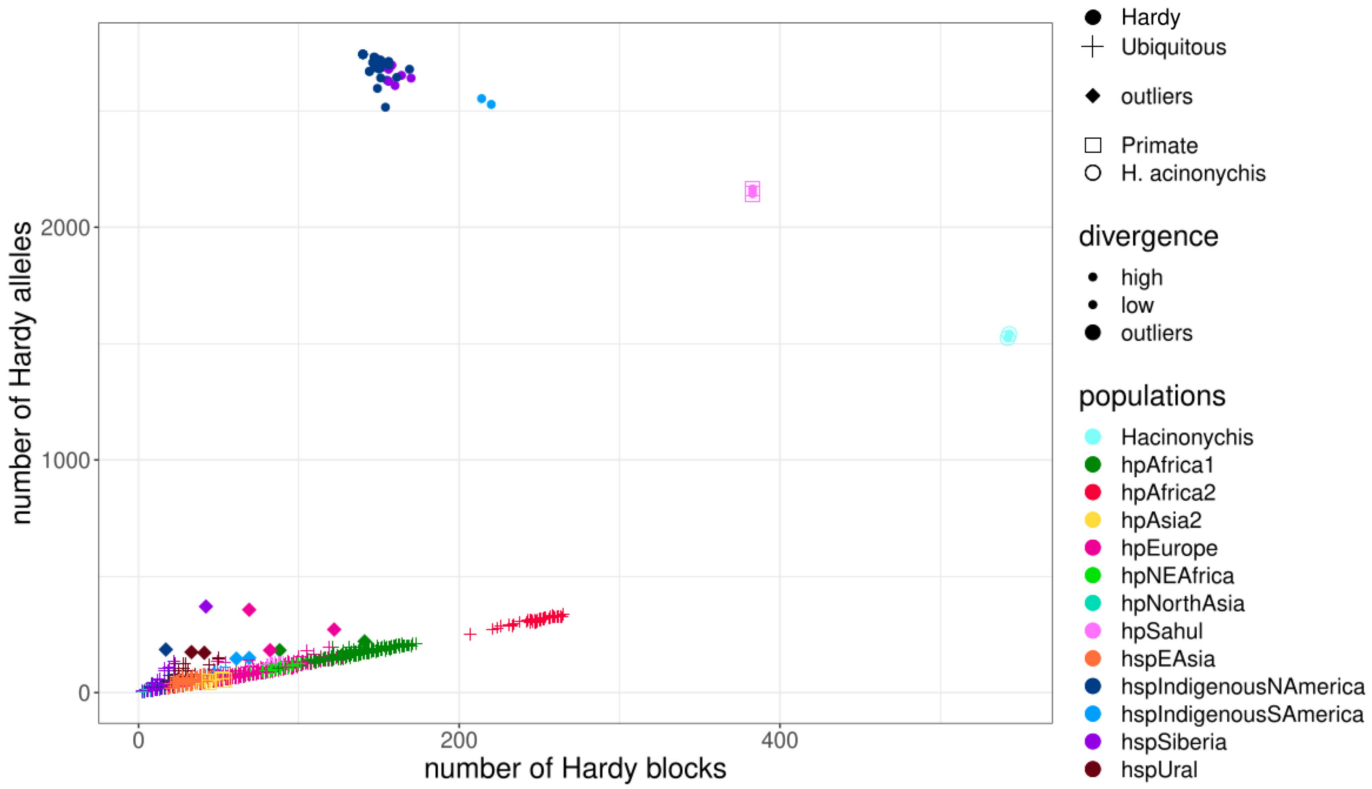


Extended Data Fig. 7 | Dot plot comparisons between genomes within and between ecospecies. The genomes of two hspIndigenousNAmerica, one Hardy and one Ubiquitous strain were plotted against the genome of strains more or less distantly related, from left to right: hspIndigenousNAmerica, hspIndigenousSAmerica, hspSiberia, hpAsia2, hpEurope, hpAfrica1, hpAfrica2, *H. acinonychis* and *H. cetorum;* and from top to bottom: Hardy vs Hardy strains, Hardy vs Ubiquitous strains and Ubiquitous vs Ubiquitous strains. Comparison between identical genomes would give single diagonal line, with breaks indicating rearrangements and differences in genome content. The presence of several small lines indicates that there are many rearrangements between the two genomes being compared. On the contrary, comparisons with long lines means highly similar genomes. For more details on how the comparisons were made, see the paragraph "Genome structure comparison" in the Method section.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Pairwise dN/dS values to relevant outgroups. dN/dS vs dS (A,B) dN/dS vs dS between the different populations (hpAfrica2 excluded) and hpAfrica2 for the undifferentiated (A) and differentiated (B) genes. Thus, all comparisons involve hpAfrica2 strains and the dots are coloured based on the non-hpAfrica2 population. In addition, the shape represents the ecospecies of the non-hpAfrica2 strain, the dots represent Hardy strains while the crosses represent the Ubiquitous strains; the primates and *H. acinonychis* strains are indicated with squares and circles, respectively. (C,D) dN/dS vs dS between the Ubiquitous and Hardy strains for the undifferentiated (C) and differentiated (D) genes (subplots based on whether the Hardy strains were *H. acinonychis* or non-*H. acinonychis*). For the C and D subplots, all comparisons involve one Hardy (*H. acinonychis* or non-*H. acinonychis*) and one Ubiquitous strain, and the dots are coloured based on the Ubiquitous strain population. In all cases, each dot represents the value for a non-outgroup strain, averaged over their values when compared against the different outgroup strains (the outgroups are hpAfrica2 for subplots A and B and *H. acinonychis* or Hardy *H. pylori* for subplots C and D). (E) Relationship between dS and Average Nucleotide Identity (ANI) for the same comparisons as shown in panel D.



Extended Data Fig. 9 | Number of Hardy alleles per strain against the number of Hardy blocks. The dots represent the Hardy strains, and the points are coloured based on their population. The outliers from Fig. 3b are shown with a diamond. The primate and *H. acinonychis* strains are indicated by squares and circles, respectively.

nature portfolio

Corresponding author(s): Daniel Falush

Last updated by author(s): July 22, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No softwares were used for the data collection
Data analysis	Genome assembly: SPAdes v3.15.3, BACTpipe pipeline v3.1.0 Filtering: Kraken v2.1.2 Core genome variant calling: MUMmer v3.20 Alignment to the reference genome: nucmer v3.1 Calling the variant from the whole genome alignment: snp-sites v2.5.2 Extraction variants VCFtools v0.1.17 Population assignment: fineSTRUCTURE, ChromoPainter v2 PCA: PLINK v1.9 ML trees: FastTree v2.1.10 NJ trees and rooting: R package ape v5.7-1 Genes alignment: MAFFT v7.505 GWAS: R package bugwas v0.0.0.9000 (use GEMMA v0.93) FST: R package PopGenome v2.7.5 Pangenome analysis: prokka v1.4.6, Panaroo v1.2.8, Pheatmap v1.0.12 BLAST H. cetorum comparison: blastn v2.11.0 Genome structure comparison (dotplots): Gepard v1.40 Functional enrichment analysis: website DAVID 2021 dN/dS: PAML (YN00) v4.9 ANI: FastANI v1.34

Global analysis and plotting: R v4.3.1 (and R packages ggplot2 v3.3.6, ggtree v3.2.1, tidyverse v1.3.2), python v3.10.6 (and library numpy v1.23.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Code Availability Statement

The scripts for the following analysis: PCA, phylogenetic analysis, GWAS, FST and dN/dS are available on github (https://github.com/EliseTourrette/HpEcospecies_Tourrette2023, doi: 10.5281/zenodo.12740447).

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The list of strains used are provided in the Supplementary Data Tables 1 and 2, with NCBI accession numbers for all newly sequenced strains in Supplementary Data Table 1. The full dataset is also available on the Enterobase worksheet https://enterobase.warwick.ac.uk/a/108555.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No report on sex and gender of the human host.
Reporting on race, ethnicity, or other socially relevant groupings	The host ethnicity of the indigenous strains was provided when available from already published information (public isolates or isolates already published as MLST data).
Population characteristics	The clinical samples were collected in several different cohorts over the last few decades and were selected to represent geographical areas or human populations. They are not necessarily a representative sample either geographically or pathologically since H. pylori requires endoscopy, an invasive medical intervention and therefore are collected opportunistically, normally from middle age people with some kind of gastric complaint.
Recruitment	Since H. pylori requires endoscopy, an invasive medical intervention, the samples were collected opportunistically, normally from middle-aged people with some kind of gastric complaint.
Ethics oversight	Ethical permission for the collection of human gastric biopsy material had been obtained for all cohorts, including informed consent from the participating individuals. For details on the board/committee and institution that approved each study protocol, see the Ethics and Inclusion statement: Umeå University ethics committee; study conducted in accordance with the Helsinki declaration. Institutional Ethical Review Committee of the Research Institute for Gastroenterology and Liver Diseases at Shahid Beheshti University of Medical Sciences, Tehran, Iran (IR.SBMU.RIGLD.REC.1395.878, RIGLD 722, RIGLD 878). Ethics committee of the Oita University Faculty of Medicine (No. P-10-12), Japan.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*

Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Based on a global dataset of 8186 Helicobacter whole genome sequences, we define a new ecospecies found in indigenous human populations (North America and Siberia).
Research sample	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 (as of Jun 6th 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; a large number (920+190) of worldwide DNA samples were contributed by Prof Mark Achtman. Lastly, 266 strains were from Iran, collected by Abbas Yadegar, and 133 genomes from different parts of the world, including 89 from the Swedish Kalixanda cohort.
Sampling strategy	The cohorts, sampling procedure and bacterial isolation is detailed in the Methods section. The sampling of the new Siberian isolates has been detailed previously, for example in Moodley et al. 2021 (MLST).
Data collection	New sequences were collected and sequenced in four different centers: Oita University, University of Warwick, Karolinska Institutet and University of Gothenburg.
Timing and spatial scale	The clinical samples were collected in several different cohorts over the last few decades and were selected to represent geographical areas or human populations rather than reflecting a specific time interval. They are not necessarily a representative sample either geographically or pathologically since H. pylori requires endoscopy, an invasive medical intervention and therefore are collected opportunistically, normally from middle age people with some kind of gastric complaint.
Data exclusions	Excluded genomes: redundant genomes (sequences with less than 200 bp SNP distance) and low quality genomes based on assembly fragmentation (>500 contigs), coverage to the 26695 H. pylori reference strain (<70%) and contamination (<90% H. pylori). Final dataset used in the analysis: 6626 genomes see Methods section (Genome collection and Table TS1)
Reproducibility	Multiple isolates from the same geographical zone/population were sampled in order to assess the variability of the different measurements.

Randomization	The group allocations, apart from geographical origin, were the H. pylori populations, which were inferred from the analysis results (fineSTRUCTURE analysis, Supp Figure S3)
Blinding	As the categories in terms of population assignment were central to the downstream analyses and result interpretation, blinding of the group allocation was not suitable for this study.

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the*

Specimen provenance	*issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.