Repeated out-of-Africa expansions of *Helicobacter pylori* driven by replacement of deleterious mutations

Harry A. Thorpe[1&], Elise Tourrette[2&], Koji Yahara[3&], Filipa Vale[4], Siqi Liu[5], Mónica Oleastro[6], Teresa Alarcon[7], TsachiTsadok Perets[8], Saeid Latifi-Navid[9], Yoshio Yamaoka[10], Beatriz Martinez-Gonzalez[11], Ioannis Karayiannis[11], Timokratis Karamitros[11], Dionyssios N. Sgouras[11], Wael Elamin[12,13], Ben Pascoe[14], Samuel K. Sheppard[14], Jukka Ronkainen[15,16], Pertti Aro[17], Lars Engstrand[18], Lars Agreus[19], Sebastian Suerbaum[20, 21, 22§], Kaisa Thorell[23, 24§], Daniel Falush[5*§]

1 Institute of Basic Medical Sciences, University of Oslo, Oslo, harry.thorpe@medisin.uio.no
2 3 rue du Dr Lauriat, 91400 Orsay, France. elise.tourrette@gmail.com
3 Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Toyo, Japan, k-yahara@nih.go.jp
4 Pathogen Genome Bioinformatics and Computational Biology, Research Institute for Medicines (iMed-ULisboa), Faculty of Pharmacy, Universidade de Lisboa, Lisboa, Portugal vale.filipa@gmail.com
5 Centre for Microbes Development and Health, Institute Pasteur Shanghai. sqliu@ips.ac.cn, danielfalush@googlemail.com
6 National Reference Laboratory for Gastrointestinal Infections, Department of Infectious Diseases, National Institute of Health Dr Ricardo Jorge, Lisboa, Portugal. Monica.Oleastro@insa.min-saude.pt
7 Department of Microbiology, Hospital Universitario La Princesa, Instituto de Investigación Sanitaria Princesa, Madrid, Spain, talarcon@helicobacterspain.com
8 Division of Gastroenterology, Rabin Medical Center, Petah Tikva, Israel, tzahipe1@clalit.org.il
9 Department of Biology, Faculty of Sciences, University of Mohaghegh Ardabili, Iran, s_latifi@uma.ac.ir
10 Department of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Yufu, Oita, Japan yyamaoka@oita-u.ac.jp
11 Laboratory of Medical Microbiology, Hellenic Pasteur Institute, Greece, bmartinez@pasteur.gr, y.karayiannis@hotmail.com, tkaram@pasteur.gr, sgouras@pasteur.gr
12 G42 Healthcare; Abu Dhabi, wfmelamin@yahoo.com
13 Elrazi University; Khartoum; Sudan
14 The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK, bp467@bath.ac.uk, sks70@bath.ac.uk
15 Center for Life Course Health Research, University of Oulu;
16 Primary Health Care Center, Tornio, Finland jukka.ronkainen@fimnet.fi
17 Arokero Oy, Tornio, Finland. pertti.aro@fimnet.fi
18 Center for Translational Microbiome Research, Department for Microbiology, Tumor, and Cell Biology, Karolinska Institutet, Stockholm, Sweden, lars.engstrand@ki.se
19 Division of Family Medicine, Karolinska Institutet, Stockholm, Sweden, lars.agreus@ki.se

20 Department of Medical Microbiology and Hospital Epidemiology, Max von Pettenkofer Institute, Faculty of Medicine, LMU Munich, Munich, Germany, suerbaum@mvp.lmu.de
21 Department of Medical Microbiology and Hospital Epidemiology, Hannover Medical School, Hannover, Germany
22 DZIF German Center for Infection Research, Hannover-Braunschweig and Munich Partner Sites, Germany
23 Institute of Biomedicine, Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden. Kaisa.thorell@gu.se
24 Department of Clinical Microbiology, Sahlgrenska University Hospital, Gothenburg, Sweden
& contributed equally.
§ Shared senior authors
* Corresponding author: Daniel Falush, danielfalush@googlemail.com

Abstract

All genomes mutate but the consequences of the resulting deleterious mutational load are poorly understood. *Helicobacter pylori* lives in the human stomach, has a higher mutation rate than most bacteria and has accompanied anatomically modern humans in migrations including the out-of-Africa expansion more than 50,000 years ago. *H. pylori* from East Asia have accumulated at least 500 more non-synonymous mutations than African strains, which we propose is due to reduced efficacy of selection during the out-of-Africa bottleneck. *H. pylori* from Europe and the Middle East trace a substantially higher fraction of ancestry from modern African populations than the humans that carry them, which we find is due to at least three separate admixture events. African ancestry is elevated at positions in the genome where non-synonymous mutations are at high frequency in Asia. We propose that this is due to replacement of deleterious mutations that accumulated during the bottleneck, with the high overall African ancestry proportion due to clonal expansion of strains of African origin. We use simulations to show that a Muller's ratchet like effect can lead to long-term segregation of deleterious mutations within bacterial populations after a bottleneck, despite high rates of homologous recombination, but that population fitness can be restored by migration of small numbers of bacteria from non-bottlenecked populations. Our results demonstrate that population bottlenecks can have long-term genomic and demographic consequences, even in species with enormous population sizes.

Main text

Introduction

Helicobacter pylori is the dominant bacterial member of the human stomach microbiota in infected individuals and is an aetiological agent in most cases of gastric cancer, gastric lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma) and gastroduodenal ulcer disease (*1*). *H. pylori* causes chronic, decades-long infections and is often acquired within the household, limiting the rate of its diffusion through human populations in comparison with more readily transmissible pathogens (*2*). Genetic variation in bacterial genome sequences shows a phylogeographic pattern similar to that of its host, consistent with an inference that human and bacterial genes are often spread by the same migrations (*3-5*). However, the *H. pylori* population found in Europe and other parts of Eurasia is admixed, with many strains having more than half of their DNA attributable to populations closely related to those prevalent in Africa(*3, 6-8*). There is evidence for several recent human migrations out of Africa (*9*), but together they have only contributed a small fraction of the ancestry of non-Africans. This discrepancy in ancestry proportions between the bacteria and their hosts implies that African *H. pylori* has been spread to Eurasia by movements of people that have left weaker signals in human DNA.

To understand why African *H. pylori* have contributed extensive ancestry within parts of Eurasia, we have assembled a collection of strains from Europe and the Middle East, from putative source populations in Africa as well as less-admixed strains in Asia. We infer a

recent demographic history of the European and Middle Eastern strains that includes genetic drift, migration and admixture from external sources. We show that there have been at least three admixture events from African source populations that have each contributed substantial ancestry. By examining the distribution of non-synonymous mutations in different populations, we conclude that there was a large and lasting increase in the frequency of segregating deleterious mutations resulting from the out-of-Africa bottleneck. When African strains reached Eurasia due to later contact between humans, they, and the DNA they contained, had a fitness advantage and were able to spread. In the process, they reduced the mutational load in the newly admixed populations.

Repeated African Admixture into Europe and the Middle East

Based on fineSTRUCTURE clustering (Figure S1), we grouped European and Middle Eastern strains into four subpopulations named hspEuropeNEurope, hspEuropeCEurope, hspEuropeSWEurope and hspEuropeMiddleEast according to the locations they were most commonly isolated from. The first Europe in the name indicates they are subpopulations of hpEurope but for brevity we omit this part of the name in the rest of the manuscript. We investigated their sources of external ancestry by performing *in silico* chromosome painting (Figure 1A), using donor strains from three African populations (hspENEAfrica, hspCNEAfrica, hspAfrica1WAfrica) and two Asian populations (hpAsia2 and hspEAsia). Representative strains from these populations were selected from a larger collection on the basis that they showed little sign of recent admixture from other continents (Methods). The chromosome painting analysis supports previous findings that there is a North-South cline in the overall proportion of African ancestry and that hpAsia2 is a closer relative of the pre-admixture population than hspEAsia (*3, 7*). However, all isolates are painted with a substantial and largely consistent fraction of hspEAsia (with the hpAsia2:hspEAsia ratio varying from 1:1.74 for hspNEurope to 1:2.02 for hspMiddleEast), implying that hpAsia2 is not a close surrogate for the pre-admixture population.

To investigate the demographic history of admixture further, we measured genetic drift profiles separately for each ancestry component. Specifically, we compared the painting profile of pairs of admixed individuals to identify regions of the genome in which they were painted by the same donor population. For these genomic regions, we recorded if they were painted by the same specific donor strain (Methods). High rates of painting by the same donor strain indicates shared genetic drift within that ancestry component.
Genetic drift profiles for the hpAsia2 and hspEAsia ancestry components show a similar pattern across the four hpEurope subpopulations, as indicated by near identical pattern of colours for these two ancestry components in Figure 1B. The indistinguishable drift profiles provide evidence that within each hpEurope subpopulation, both components have been affected similarly by genetic drift. The simplest explanation is that both components are both being used to paint a single ancestry source that persisted in western Eurasia since the

out-of-Africa bottleneck. Therefore, the data does not provide evidence for either ancient or more modern genetic contributions from the East.

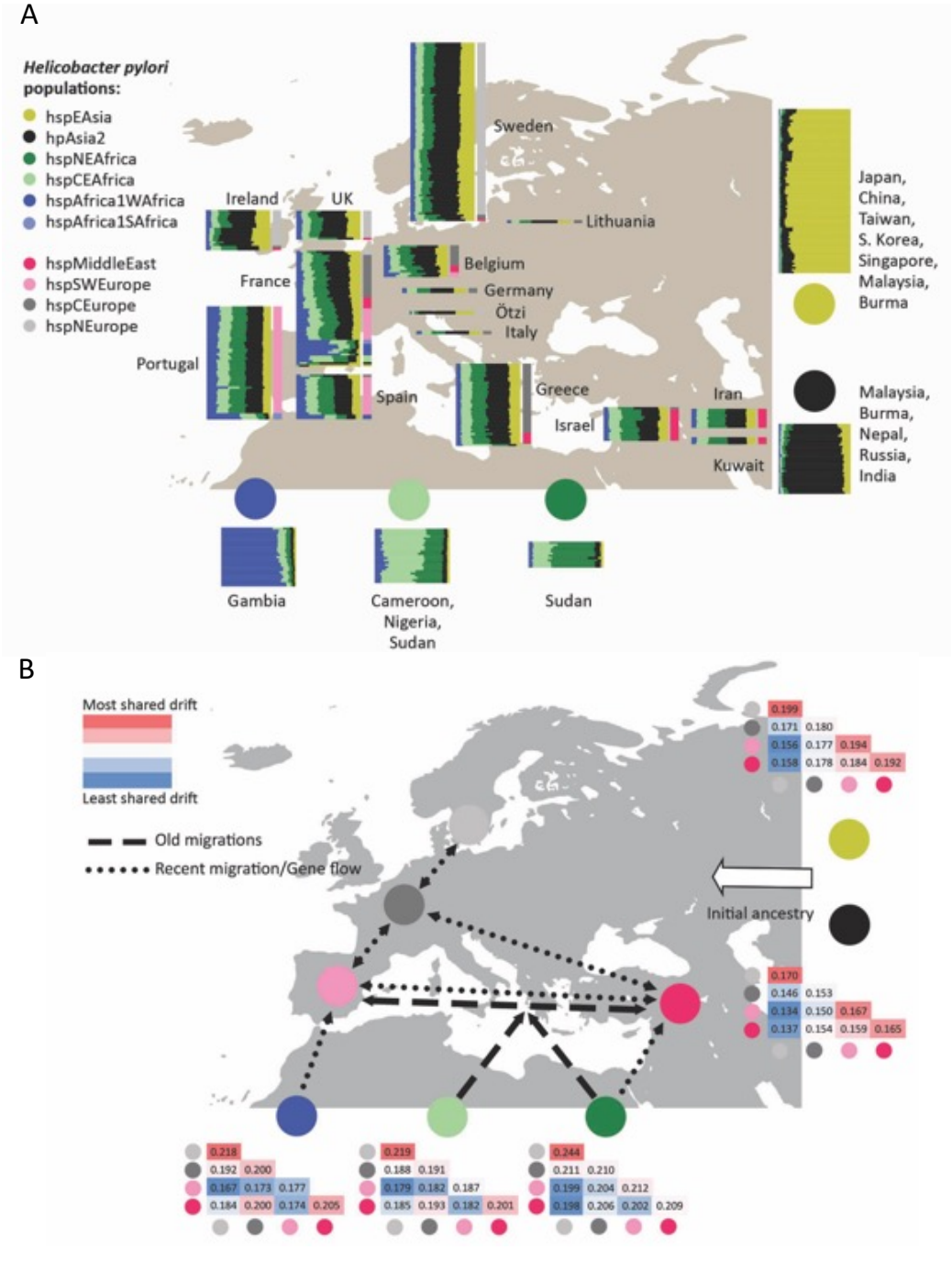

Figure 1 *Ancestry and migration history of hpEurope isolates. (A) Painting profiles of hpEurope isolates and their putative ancestral populations from Africa and Asia showing proportion of each genome(horizontal bar) painted by each of 5 ancestral donor populations (circles). HpEurope isolates are grouped by country of isolation, with bars to the right indicating the* H. pylori *population each strain is assigned to. The representative isolates from each donor population are grouped by population, with country of isolation for each group. (B) Genetic drift profiles for hpEurope subpopulations and inferred migration history.*

In addition to the North-South cline, there is also an East-West ancestry cline in the source of African admixture (Figure 1A), with distinct drift patterns for each African component seen in the four hpEurope subpopulations (Figure 1B). Strains from hspSWEurope have the highest fractions of hspAfrica1WAfrica and this ancestry component shows low levels of drift, implying that this subpopulation has undergone recent admixture from strains closely related to those currently found in West Africa. Furthermore, there are strains from Spain, Portugal and France assigned to hspAfrica1 subpopulations and a strain in Portugal with an intermediate ancestry profile suggesting they have arrived within the last few human generations. Strains from the hspMiddleEast subpopulation have the highest fraction from hspENEAfrica and the lowest levels of drift in this component. These two populations have therefore received genetic material from Africa subsequent to the initial gene flow that has introduced ancestry across the continent. These patterns imply that there have been at least 3 separate admixture events involving distinct populations of African bacteria.

The drift components also provide evidence about local migration. hspMiddleEast and hspSWEurope have high levels of shared drift in both Asian components, implying that prior to admixture, these two populations were closely related. hspNEurope has high subpopulation-specific drift in all ancestry components, showing that it has undergone recent genetic drift, while hspCEurope has almost none in any component, suggesting that it has been a hub for migration between populations.

The genome of a strain colonizing the Tirolean iceman, Ötzi, has been inferred to be a nearly pure representative of the pre-admixture population based on more limited data, which was interpreted as evidence that most of the admixture took place in the 5,300 years since his death (*7*). In our fineSTRUCTURE analysis, the Ötzi genome clusters with hspNEurope isolates from Ireland and Sweden with the lowest African ancestry, one of which has the same non-African ancestry proportion as Ötzi in the chromosome painting, showing that the Ötzi genome in fact falls within modern variation in ancestry proportions in Europe. We interpret this as evidence that substantial African ancestry had already been introduced into Europe when Ötzi lived but that ancestry proportions in particular locations have changed substantially in subsequent millennia.

Overall, the chromosome painting results show that, in addition to contemporary migrations that have introduced *H. pylori* with atypical profiles into countries such as Ireland, Portugal, France and Spain (Figure 1A), *H. pylori* have spread out of Africa at least three times (Figure 1B). Each of these migrations is sufficiently old that the DNA has been absorbed into the local gene pools, leading to a high degree of uniformity in ancestry profiles for most isolates in individual locations (Figure 1A). At least one of the early admixture events was shared between the four subpopulations, spanning Europe and the Middle East, and left traces in Ötzi's genome, while later ones, labelled as "recent" in Figure 1B had foci in South Europe

and the Middle East respectively. Gene flow between the regional subpopulations has affected all ancestry components but has not been sufficient to homogenize ancestry proportions across the continent.

Evidence for a role of deleterious mutations in the repeated expansion of bacteria from Africa.

The ability of *H. pylori* of African origin to spread effectively in non-African populations on multiple independent occasions is unexpected, since the resident bacteria will have had an opportunity to adapt to local conditions. One potential explanation is that deleterious mutations accumulated in genomes of strains carried by the early waves of modern humans that spread from Africa. Demographic bottlenecks associated with these migrations have been sufficient to leave an imprint on neutral genetic variation within the human genome, which indicate a reduction in effective population size in the ancestry of non-African humans around 50,000 years ago, followed by more recent expansion (*10*).

Population genetic theory (*11*) and evidence from experimental systems (*12*) has shown that demographic bottlenecks can lead to reduction in average fitness through processes such as fixation of deleterious mutations of small effect or the stochastic loss of the fittest genomes (Muller's ratchet (*13*)). However, although the out-of-Africa bottlenecks have had measurable impact on the pattern of segregating mutations within populations (*14*), there is less evidence that individuals from non-African populations have accumulated a larger burden of non-synonymous mutations when measured relative to an outgroup (*15*). *H. pylori* shows much higher rates of genetic differentiation between geographical regions (*4*) than its host (*10*), which most likely reflects transmission bottlenecks during spread from person to person as well as the expansion of fit clones. Consequently, demographic bottlenecks that have had modest fitness consequences for humans can potentially have more substantial effects on the bacteria they carry.

Our data points to a substantial mutational burden associated with the out-of-Africa bottleneck which has been ameliorated by admixture in Europe and the Middle East. First, we calculated synonymous differentiation (dS) and the ratio between non-synonymous and synonymous changes (dN/dS values) for each of the subpopulations relative to *H. acinonychis (Figure 2B-C)*, which is the most suitable available outgroup since it is thought to have evolved independently since undergoing a host-jump to big cats, around 100,000 years ago (*6*) (*16*). The hpAfrica2 population, which originated in Khoisan and has spread to other groups in Southern Africa (*6*) has a substantially lower value of dS with *H. acinonychis* than all of the other *human H. pylori* populations, reflecting a closer evolutionary relationship (Figure 2A-B, (*16*)). The higher value of dN/dS for hpAfrica2 than for the other human populations is difficult to interpret, since it might reflect evolutionary rate changes when *H. acinonychis* jumped between hosts.

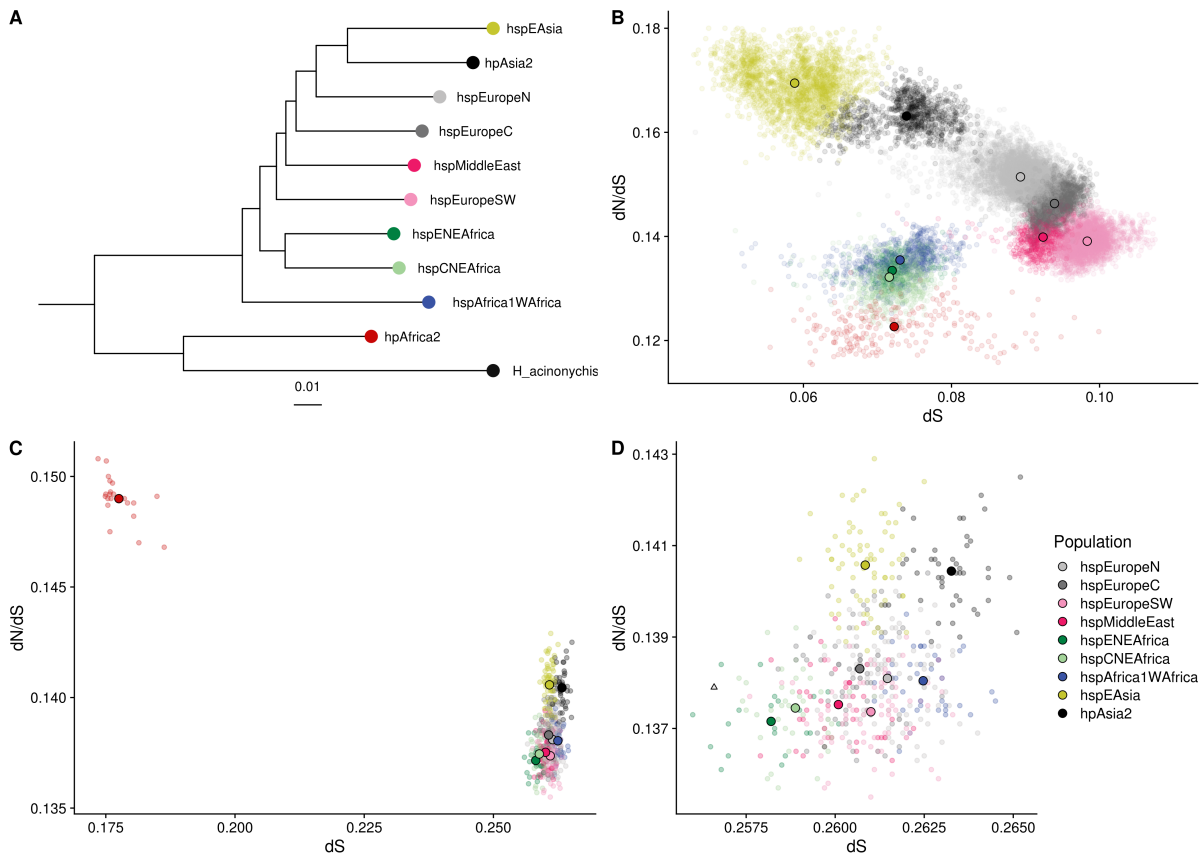

Figure 2. *Synonymous divergence (dS) and dN/dS within and between populations. (A) Neighbour joining tree calculated based on mean synonymous genetic distances between isolates in different populations. (B) dN/dS calculated to H. acinonychis outgroup (y axis) plotted against dS (x axis) for isolates (opaque points) and populations (solid points), with colours corresponding to population labels in A. (C) close up of plot B, excluding hpAfrica2 isolates. (D) within population dN/dS plotted against dS. Opaque points show pairwise distances; solid points indicate population means.*

H. acinonychis is approximately equidistant to the African and Asian populations that represent putative sources for European and Middle Eastern strains. Synonymous differentiation (dS) from *H. achinonychis* range between 0.258 for hspENEAfrica to 0.262 for hspAfrica1WAfrica, with the Asian populations having intermediate values. The source of this variation is unknown and is somewhat surprising given the small genetic distances between African subpopulations. One possibility is that the variation reflects admixture with populations such as hpAfrica2 which are closer to or further from *H. acinonychis* or alternatively historical differences in the mutation rate. In any case, there are consistent differences in dN/dS, between Africa and Asia, with values for the three African populations ranging between 0.137 and 0.138 and the Asian populations having values of 0.140 and 0.141. This corresponds to an average of 574 extra non-synonymous mutations on average in the Asian populations.

hpEurope bacteria have dN/dS values that are intermediate between African and Asian populations but are lower (0.137-0.138) than would be predicted if they were random mixtures of African and Asian genomes with the proportions estimated by chromosome painting in Figure 1 (0.139-0.140). However, this discrepancy represents weak evidence for a direct benefit of admixture, since it might instead reflect differences in dN/dS between the population that existed in Europe prior to admixture and the Asian populations that are used as surrogates for this ancestry in this analysis.

Second, we investigated genetic variation within contemporary populations (Figure 2D). Diversity levels are similar in Asian and African populations including hpAfrica2, with dS values around 0.07. However, there is a substantial difference in dN/dS values, which vary between 0.123-0.133 for African populations and 0.163-0.169 for the Asian ones. This difference corresponds to an average of 1216 non-synonymous differences between pairs of strains. European and Middle Eastern subpopulations are substantially more diverse than the others but have intermediate dN/dS values. These results show that the larger number of differences in dN/dS to the outgroup in the Asian populations reflect a larger number of segregating non-synonymous variants and thus cannot simply be attributed to fixation of mutations during the out-of-Africa bottleneck.

Third, we investigated the effect of mutations that accumulated in African and Asian populations on the ancestry of admixed bacteria in Europe and the Middle East. We hypothesized that natural selection would preferentially replace any deleterious mutations that rose to high frequency due to the out-of-Africa bottleneck. Such mutations would have been shared between Asia, Europe and the Middle East prior to admixture.
Therefore, we studied the effect of segregating synonymous and non-synonoymous mutations found in contemporary Asian populations on ancestry in Europe and the Middle East.

For each position in the alignment, we calculated a mutation score, which is the difference between African and Asian strains in the proportion of nucleotides that differed *from H. acinonychis* (with equal weight given to each subpopulation, see methods). We then investigated whether there is variation in overall Asian ancestry proportion (hpAsia2+hspEAsia in the chromosome painting) associated with this score. Most sites are non-polymorphic and have a mutation score of 0.0 which is therefore used as a baseline and compared independently to positive and negative scoring sites. To allow for correlations between adjacent sites, statistical significance of the regression was assessed using a gene-by-gene jackknife (methods).

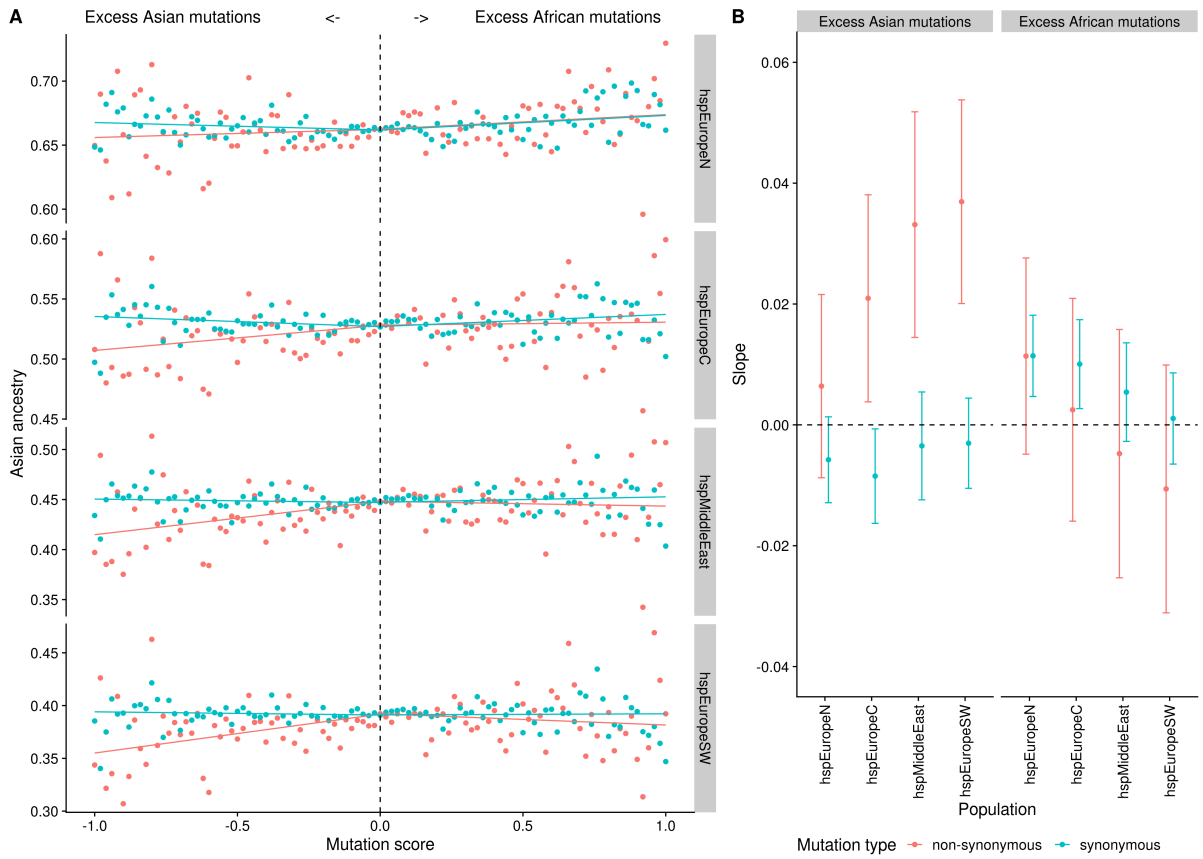

Figure 3 *Genetic ancestry of hpEurope subpopulations as a function of mutation score. (A) average ancestry in chromosome painting analyses plotted against mutation score (Average African distance to H. acinonychis vs Average Asian distance to H. acinonychis at that site) in bins of 0.02. Regression lines are calculated separately for positive and negative mutation scores (B, C) regression slopes with 95% confidence intervals estimated using a gene-by-gene jacknife for excess Asian (B), and excess African mutations (C) respectively.*

H. pylori shows little evidence for codon usage bias (*17*), so most synonymous mutations should have be neutral. Consistent with this expectation, we find that negative scores at synonymous sites, indicating derived mutations segregating at higher frequency in Asian populations, do not influence ancestry proportions in hpEurope subpopulations (Figure 3). There is a modest but statistically significant positive correlation between positive scores and Asian ancestry in two of the four subpopulations (hspNEurope and hspCEurope). This pattern is hard to interpret but might reflect differences in the spatial distribution of mutations that accumulated in African and non-African genomes.

For non-synonymous sites there is a consistent, stronger and easier to interpret pattern. A negative mutation score leads to a deficit in Asian ancestry, with the strongest regression slopes estimated for the two Southern populations, and the weakest estimated for hspNEurope. Avoidance of Asian ancestry is noticeable even for small mutation scores, implying that many of the variants found at low frequency in modern Asian populations also

segregated historically within Europe and that these polymorphisms have had a fitness disadvantage, leading to an Asian ancestry deficit at these sites. By contrast, excess mutations in African strains do not have a detectable effect on ancestry in any of the four hpEurope subpopulations, with regression slopes statistically indistinguishable from those found for synonymous positions. Thus, admixed bacteria have avoided non-synonymous mutations that accumulated in populations that have been through the out-of-Africa bottleneck, but not those that have not. We infer that a proportion of these non-synonymous mutations are deleterious and have been selected against in the admixed population.

We investigated whether specific genes were enriched for African ancestry. There is substantial variation amongst genes in the average African ancestry proportion with strong correlations between proportions in the four hpEurope subpopulations (Figure S2), which is consistent with much of the ancestry being due to a single shared admixture event. However, there is no clear differences between COG categories in average ancestry proportion (Figure S3), making it difficult to come to conclusions about selection on particular classes of genes.

Simulations recapitulate our empirical results and implicate a Muller's ratchet-like mechanism.

To gain insight into the mechanisms underlying our results, we performed simulations of bacterial populations subject to a mutational load. We specifically wanted to investigate whether bottlenecks could induce long-term reductions in fitness without extensive fixation of deleterious mutations and whether this fitness could be restored by admixture.

Muller's ratchet refers to a phenomenon where the fittest strains (i.e. with the smallest number of deleterious mutations) are stochastically lost from the population at a faster rate than new fittest strains are generated by recombination or back mutation. If the ratchet turns repeatedly then this leads to a progressive erosion of fitness. It has been shown that loss of fittest strains by Muller's ratchet is quickly reversed in populations with crossing over or independent segregation of multiple chromosomes (*18*). However, *H. pylori* only has a single chromosome, which recombines exclusively by replacement of short tracts (*19*). Therefore, it is possible that ratchet-like phenomena might operate in our simulations and in *H. pylori* populations.

We simulated evolution of populations of bacterial chromosomes of 1.6 Mbp in length exposed to neutral and deleterious mutations with selection coefficients ranging from -0.005 to -0.0001 (methods, figure 4A-4D, S4, S5). To facilitate fast simulation, recombination takes place in each bacterial genome via import of a single recombination tract at a random location in the genome. We performed simulations for tract lengths of 0 bp (corresponding to no recombination), 500 bp, 5000 bp and 50000 bp. In these simulations increasing the

lengths of the tracts, and hence the amount of genetic material exchanged between strains, progressively increases population fitness (Figure 4A). For non-recombining populations and those with tract lengths of 500 bp and 5000 bp, the variance in fitness in the population is lower than it would be if mutations were redistributed at random amongst strains in the population, but when the tracts are 50000 bp, the variance (Figure 4B) and the frequency of high fitness strains in the tail of the distribution (Figure S6) is no longer consistently lower indicating that Muller's ratchet no longer operates.

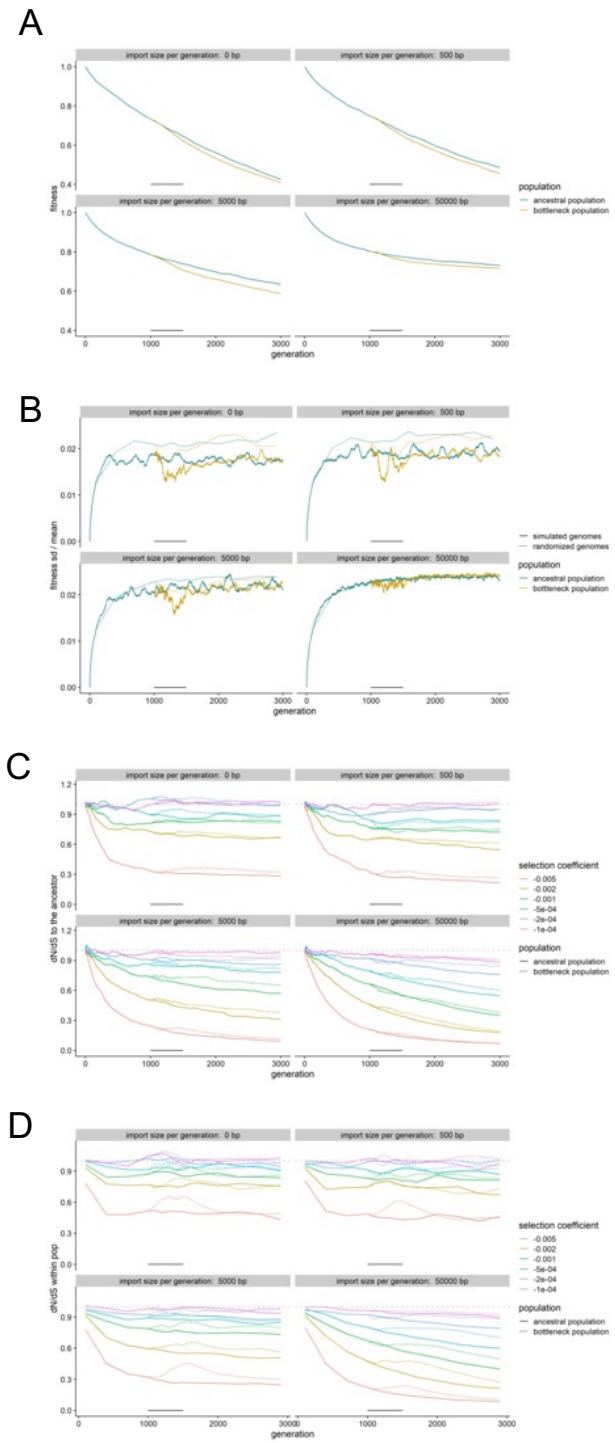

Figure 4: Simulation of the effects of a bottleneck for different level of recombination (size of the import for one generation being 0bp, 500bp, 5000bp or 50000bp for a genome of 1.6 Mbp), for the ancestral and bottleneck populations. (A) Average fitness. (B) Fitness variance, normalized by the mean fitness. Dashed lines show the effect of keeping the gene frequencies at each position the same but randomizing the mutations amongst strains (C) dN/dS to the ancestor, for the different classes of mutations. (D) Within population dN/dS, for the different classes of mutations.

In these simulations, a demographic bottleneck, consisting of a reduction in population size from 10,000 to 1,000 lasting 500 generations can lead to a long-term decrease in the mean fitness of the population, which is largest for 5,000 bp tract lengths (Figure 4A). The bottleneck also leads to an increase in dN/dS values which is most consistent and sustained in simulations with larger recombination tract lengths and for mutations with intermediate selection coefficients (i.e. 0.001 and 0.002, Figure 4C, 4D). For these combinations, the increase in dN/dS persists until the simulations finish 3,000 generations after the bottleneck, showing that bottlenecks can lead to long-term increases in dN/dS.

For the simulations with 0 bp and 500 bp recombination tracts we did not observe a similar consistent increase in dN/dS. Our interpretation is that for these recombination parameters, there is strong linkage between sites, meaning that mutations rarely escape the genetic background from which they arose and selection happens principally at the level of the clonal lineage rather than at individual sites. Consequently, genetic drift reduces diversity in synonymous and non-synonymous mutations by approximately the same proportion, in a more stochastic fashion than at higher recombination rates. These simulation results imply that elevated dN/dS values that we see in non-African populations (Figure 1A,1B) can be explained by reduced effectiveness of selection during the out of Africa bottleneck. However, they also imply that consistent effects of bottlenecks on dN/dS may not be seen in bacterial species where recombination rates are lower.

In order to investigate the effect of gene flow from Africa into Eurasian populations, we transplanted single strains from the non-bottlenecked population to the bottlenecked one at 500 generation intervals, starting in the 8000th generation of the simulations (2500 generations after the end of the bottleneck). In the simulation with 5000 bp imports, descendants of the first transplanted strain quickly rose to high frequency and eventually to fixation (Figure 5).

The transported lineage was rare initially (1/10,000), meaning that it had substantial opportunity to import genetic material to strains that had been through the bottleneck while rising in frequency. 300 generations after the transplantation event, the overall genetic contribution to the admixed gene pool from the bottlenecked population had stabilized at around 1/3 of its original value at neutral sites (Figure 5B). However,

deleterious mutations originating in the bottlenecked population fell by a larger proportion, depending on their selection coefficient. For example, at generation 8000 the average strain had 73 mutations with a selection coefficient of -0.002 that arose in the bottlenecked population but by generation 8300 this had fallen to 14. Including newly arisen mutations and those from the transplanted strain, the average genome had 70 mutations with this selection coefficient and total fitness had increased from 0.329360 at generation 8000 to 0.341028 at generation 8300. (Figure S5A).

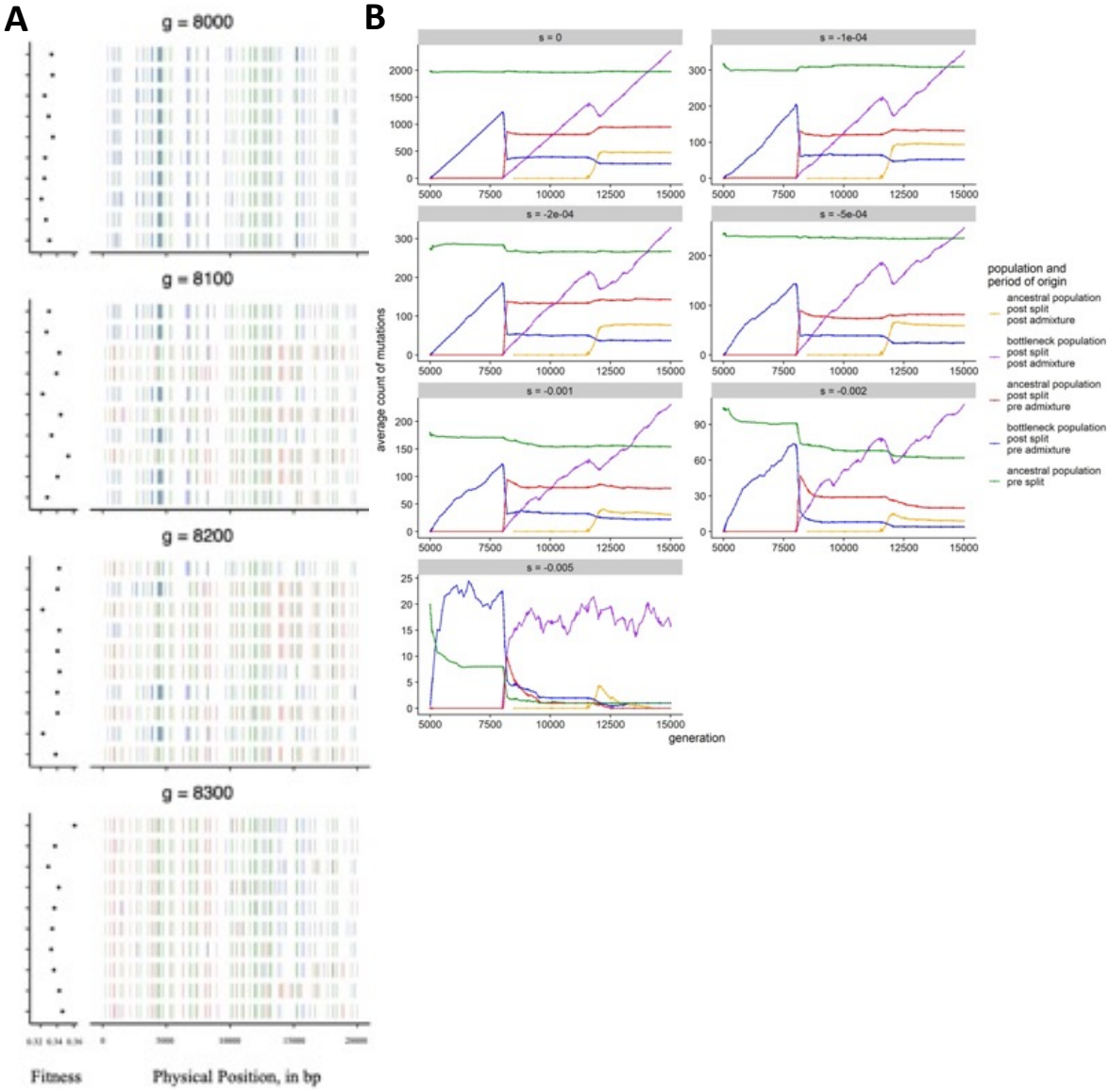

Figure 5 *Simulated admixture into bottlenecked population. The simulations presented in Figure 4, in which Population 2 underwent a bottleneck at generation 5000, were continued but starting at generation 8000, a single strain was transplanted from Population 1 into Population 2 every 500 generations. Only results for the simulation with tract length 5000bp are shown. (A) Genetic tracts from Population 2 in the aftermath of admixture. Shows first 20,000bp of 10 randomly selected chromosomes from the population at four timepoints*

after admixture began. (B) Average count of mutations in Population 2, categorized according to selection coefficient and when and where they arose.

The subsequent 5 strains that were transplanted into the bottlenecked population were lost stochastically at low frequencies, but the 6th strain (introduced at generation 11500) also rose in frequency (Figure 5B), leading to more modest infusion of genetic material from the unbottlenecked population and a small rise in average fitness. For other recombination tract lengths, none of the introduced strains left a substantial genetic contribution (Figure S5, S7). These simulations show that admixture events involving small numbers of migrant strains can contribute substantial amounts of genetic material, with a disproportionate genetic contribution at sites where deleterious mutations segregate in the bottlenecked population, consistent with our results from admixed *H. pylori* populations (Figure 3).

Discussion

Our simulations of recombining bacterial chromosomes undergoing demographic bottlenecks and admixture events have replicated several features of the *H. pylori* data presented here, showing that deleterious mutations provide a compelling explanation for the repeated spread of African *H. pylori* into other continents. First, our simulations replicate the elevated dN/dS values in populations that have been through demographic bottlenecks, as found in East Asian and other non-admixed *H. pylori* populations. Secondly, we show that populations that have been through bottlenecks are susceptible to repeatedly being invaded by bacteria from non-bottlenecked populations. Thirdly, we have shown that the resulting admixture events can lead to highly mosaic genomes, as observed in Europe and elsewhere. Fourthly, we show that ancestry from invading lineages is higher in regions of the genome where non-synonomous mutations are segregating within the bottlenecked populations, implying that on average, these mutations are deleterious.

While the deleterious mutation hypothesis provides an excellent qualitative explanation for the patterns of mixture that we have described, there are several facets that require elaboration. We have shown that tiny numbers of migrant bacteria are potentially able to substantially alter the genetic makeup of populations. However, little is known about the nature of contacts between African and non-African populations in prehistory. Furthermore, the spread of bacteria once they left Africa would have depended on the rate at which *H. pylori* is transmitted between individuals from different households, which is poorly characterized, despite evidence of mixture of *H. pylori* of different origins within extended families in contemporary settings (*20*).

There are large differences in African ancestry proportion between European countries, implying that later admixture events have been sufficiently recent that mixture frequencies have not been homogenized by genetic exchange within the continent. Furthermore, the single ancient genome from Ötzi's stomach is more similar in ancestry profile to

contemporary bacteria from Ireland than from Southern Europe where he lived, suggesting that the 5,300 year-old genome had ancestry from the earliest waves of bacterial DNA from Africa, but not from the recent ones responsible for the modern East-West cline.

An intriguing question is why some populations appear to have been resistant to invasion by African *H. pylori*. For example, within East Asia, most strains appear to come from the hpEastAsia population, with very little evidence of admixture. Lack of contact with Africans does not seem a sufficient explanation, given the large number of documented contacts between East Asians and other Eurasian populations within the last several thousand years. The high burden of *H. pylori* related gastric disease in the region is notorious and hpEastAsia bacteria are known for distinctive variants at virulence associated loci including *cagA* and *vacA* (*21*). It is possible that strains from this population have acquired a suite of adaptations that allows them to outcompete invading bacteria despite the large mutation load within their genomes. This raises the possibility that some of the non-synonomous mutations that rose to high frequency during the bottleneck may have allowed rapid adaptation, in other words a form of evolution by shifting balance (*22*).

H. pylori seems to be an outlier amongst bacteria in many features of its biology including its slow rate of spread between human populations and its high mutation and recombination rates. Our results suggest that recombination may save strains from rapid mutational meltdown but that deleterious mutations persist within populations, with the effects of bottlenecks enduring for millennia. The unusual properties of *H. pylori* make it a powerful model system for understanding how deleterious mutations interact with demographic processes and adaptive ones to mould diversity within natural populations.

Methods

Dataset collection

A dataset of 716 *Helicobacter pylori* whole-genome sequences was assembled consisting of 213 newly sequenced isolates from Europe, Asia and Africa (supplementary table 1) and selected publicly available genomes (supplementary table 2).

Genome sequencing

Genomes new for this study were sequenced at five different centres, Karolinska Institute, Sweden (KI), Hannover Medical School, Hannover, Germany (MHH), Hellenic Pasteur Institute, Greece (HPI), Oita University, Japan (OiU), and University of Bath, UK (UBa), see supplementary table 1.

Genomic DNA from strains marked with KI in supplementary table 1 was extracted using DNeasy Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's guidelines for Gram-negative bacteria. Sequencing libraries were prepared using the TruSeq Nano kit (Illumina, San Diego, CA, USA) and sequenced on the MiSeq platform, v3 chemistry, using 300 bp paired end mode.

Genomic DNA from strains marked MHH was isolated from *Helicobacter pylori* strains after 24h culture on *Helicobacter pylori* selective agar (in-house recipe) with the Genomic Tip 100/G (Qiagen, Hilden, Germany). Nextera XT libraries were generated and sequenced in three different runs on MiSeq 2x300bp paired (Illumina, San Diego, CA, USA), as recommended by the manufacturer. All quantification steps of gDNA and NGS libraries were done with Qubit dsDNA HS Assay Kit (Invitrogen, ThermoFisher Scientific, Carlsbad, CA, USA).

For strains marked with HPI adapter-compatible DNA was prepared using Ion Xpress™ Plus Fragment Library Kit and enzymatically fragmented for 5-12 minutes, resulting in a median fragment size of 350-450 bp and the libraries were prepared using the Ion Plus Fragment Library Kit. The resulting 400 bp insert libraries were used to prepare enriched, template-positive Ion PGM™ Hi-Q™ View Ion Sphere Particles (ISPs) with the Ion OneTouch™ 2 System. 850-flows sequencing was performed using the Ion PGM™ Hi-Q™ View Sequencing Kit with the Ion 318™ Chip Kit v2.

For genomes marked with UBa, genomic DNA was quantified using a NanoDrop spectrophotometer, as well as the Quant-iT DNA Assay Kit (Life Technologies, Paisley, UK) before sequencing. High-throughput genome sequencing was performed using a HiSeq 2500 machine (Illumina, San Diego, CA, USA).

Genomic DNA from strains marked with OiU was extracted using DNeasy Blood & Tissue kit (QIAGEN, Hilden, Germany). DNA concentration was measured using QuantusTM Fluorometer (Promega). High-throughput genome sequencing was performed either on Hiseq 2000 (2 × 100 or 2 × 150 paired-end reads) or Miseq (2 × 300 paired-end reads) sequencer (Illumina, San Diego, CA) following the manufacturer ‘s instruction.

Primary bioinformatics analysis

For the KI genomes the raw sequencing reads were quality trimmed and filtered using TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) applying a minimum q-score of 30, and *de novo* assembled using SPAdes (*23*) with the –careful option. Contigs with very low coverage and that were shorter than 500 bp were discarded prior to annotation.

For the MHH genomes quality filtering was done with Trimmomatic version 0.36 (*24*) and the assemblies were performed with SPAdes genome assembler v. 3.9.0 and resulted in number of all contigs from 26 up to 117. Assemblies were quality controlled with QUAST (*25*).

For the HPI genomes quality control of raw sequencing reads was performed using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Unbiased *de novo* assembly was performed using SPAdes genome assembler v.3.5.0in default mode.

For the UBa genomes, raw sequencing reads were quality trimmed and filtered using Trimmomatic version 0.33 and the 100-bp short read paired-end data was assembled using the *de novo* assembly algorithm Velvet version 1.2.08 (*26*). The VelvetOptimiser script (version 2.2.4) was run for all odd k-mer values from 21 to 99. The minimum output contig size was set to 200 bp with default settings, and the scaffolding option was disabled.

For the OiU genomes, Trimmomatic v0.35 was used to remove adapter sequences and low-quality bases from raw short reads data. Trimmed reads were then de novo assembled to produce contigs using SPAdes genome assembler v3.12.0 with the "-careful" option to reduce mismatches in the assembly. The minimum contig length was set to 200 bp

Annotation of both newly sequenced draft genomes and publicly available sequences was performed using the prokka annotation pipeline v 1.12 (*27*) using the most recent version of the 26695 genome (*28*) as primary annotation source.
Genome size and contig/scaffold number was collected from the prokka annotation output using the MultiQC tool (*29*) and collected into Supplementary table 3. All newly sequenced genomes were submitted to GenBank under BioProject PRJNA479414.
All strains, their population designations and their role in the respective analyses is shown in Supplementary table 4.

Sequence comparison and alignment:
All isolates were mapped to the 26695 genome (NC000915.1) using the Snippy software version 3.2-dev (https://github.com/tseemann/snippy). The resulting core genome, which was collected with the same tool, contained 287 746 core SNPs from 979 771 variant sites.

FineSTRUCTURE
We inferred population structure among the strains based on the genome-wide haplotype data of the reference-based alignment to 26695 described above, using the chromosome painting and fineSTRUCTURE (*30*) according to a procedure of our preceding study that applied them to *H. pylori* genome(*31*). Briefly, we used ChromoPainter (version 0.04) to infer chunks of DNA donated from a donor to a recipient for each recipient haplotype and summarized the results into a "co-ancestry matrix" which contains the number of recombination-derived chunks from each donor to each recipient individual. We then ran fineSTRUCTURE (version 0.02) for 100,000 iterations of both the burn-in and Markov chain Monte Carlo (MCMC) chain in order to conduct clustering of individuals based on the co-ancestry matrix.

Choice of donor and recipient strains and chromosome painting

D-statistics were calculated for strains assigned to each of the five ancestral populations hspEAsia, hpAsia2, hspAfrica1CNEAfrica, hspAfrica1ENEAfrica and hspAfrica1WAfrica. D-statistics were calculated using popstats (https://github.com/pontussk/popstats) and specifying individual A as SouthAfrica7 (hpAfrica2), individual B as GAM260Bi (hspAfrica1WAfrica), individual Y as F227 (hspEAsia) and individual X. In this comparison, negative D-statistics imply more African ancestry in the strain designated as individual X than in F227. D-statistics values can be found in Supplementary Table 5 and Supplementary Table 6.

A subset of strains from each of the five ancestral populations were chosen as donors to get groups of similar size and were selected based on the fineSTRUCTURE analysis to get good representativeness over the donor populations. Some hpAsia2 strains showed signs of elevated African admixture based on negative D-statistics values and these strains were not selected as donors.

We conducted chromosome painting of 646 recipient strains (belonging to hspNEurope, hspCEurope, hspSWEurope, hspMiddleEast, hspENEAfrica, hspEAsia, hspCNEAfrica, hspAfrica1WAfrica, hpAsia2, and hpAfrica2). For this purpose, we used ChromoPainterV2 software (*26*). For each recipient population, we calculated site-by-site average copying probability from each of the five donor populations. Gene-by-gene averages were also calculated by averaging of the sites in each gene for the 790 genes in the alignment. Regressions between gene-by-gene averages in different hpEurope subpopulation were calculated using the aq.plot() function in R.

Shared drift estimation

We used the chromosome painting analysis to investigate shared genetic drift profiles, within and between hpEurope subpopulations with different ancestry components. For each donor population used in the painting we tabulated the number of sites in the genome that each pair of recipient individuals used the same donor individual in the painting as well as the proportion of sites in which they copied from the same donor population (same individual or different individual). These two numbers were summed up for recipient individuals from particular combinations of hpEurope subpopulation pairs with the ratio giving the estimate of shared drift.

dN/dS calculations

From the fineSTRUCTURE analysis, a sub-dataset was collected consisting of the European strains assigned to the hpEurope populations, together with a representative selection of strains from the ancestral populations hpAfrica2, hspAfrica1WAfrica, hspAfrica1CNEAfrica, hspAfrica1ENEAfrica, hpAsia2, and hspEAsia. The *H. acinonychis* genome was added to this

dataset to provide an outgroup. For a detailed list of which strains that were included in these analyses, see columns in table supplementary tables 1 and 2.dN/dS was estimated pairwise between these strains from core genome alignments using the method of Yang and Nielson (*32*), as implemented in Paml v4.7.

To calculate the numbers of excess mutations observed in the Asian populations, the mean number of non-synonymous mutations was calculated for African and Asian populations, and the difference was then multiplied by a correction factor of 1.37 to account for the loss of some coding sites in the core genome alignment (1.11Mb) compared with the reference genome (1.52Mb).

Ancestry and mutation score analysis

Using *H. acinonychis* as a reference, non-synonymous and synonymous mutations were called against representative strains from the African and Asian populations. For this analysis the ancestral populations were combined to reduce the ancestry components to either African (hpAfrica) or Asian (hpAsia). For each site, the frequencies of these mutations within each ancestral population were calculated to give a score between 0 and 1. These scores were then combined by subtracting the hpAsia scores from the hpAfrica scores to give a score between -1 and 1, where -1 = fixed in hpAsia, absent in hpAfrica, 0 = equal frequencies in hpAsia and hpAfrica, 1 = absent in hpAsia, fixed in hpAfrica. Thus, these scores represent the mutational load for each site in the ancestral populations. We then combined these site by site scores with the site by site chromosome painting data for the European populations so that each site had an estimate of ancestry component from both Asia and Africa, and a mutational score representing the mutational load in the ancestral populations.

We then computed linear regressions of hpAsian ancestry against mutation score for each European population, and we did this for the positive (excess African) and negative (excess Asian) mutation scores separately. To confirm that the regression slopes were not driven by a small number of outlier genes we conducted a gene-by-gene jackknife by repeating the regressions but removing all the sites from a single gene each time. From this distribution of slope values we calculated the pseudovalues as pseudovalue=slope-((n-1)*(j_slope-slope)). We then calculated the mean and confidence limits of these pseudovalues as mean(pseudovalues) ± quantile(p=0.05/2, df=n-1) * se(pseudovalues).All data manipulation and statistical analysis was performed in R 3.6.1. Packages from the 'tidyverse' collection were used extensively, and ggplot2 was used for plotting.

Simulation of bacterial populations evolving under deleterious mutation pressure.

We used SLiM v3 (*33*) to simulate the accumulation of deleterious mutations in bacterial populations. Each bacterial genome was a single haploid chromosome of length 1.6 Mb, with a mutation rate per site of $5x10^{-7}$. Half of the mutations are neutral, with the other half

being deleterious with six different selection coefficients (-5e-3, -2e-3, -1e-3, -5e-4,-2e-4, -1e-4), each accounting for 1/12 of the mutations. There is no back mutation. Mutations have a multiplicative effect on fitness.
At the beginning of the simulation, there is no variation in the population. A single population of size 10,000 evolves for 1000 generations. 1000 of the strains are moved into a second population, which retains this size for 500 generations before expanding to size 10,000. The simulations are completed after 3000 total generations. Once every 100 generations, the full genome (mutations present in each strain, their position, selection coefficient) is saved. Every generation, the number of mutations per strain, the fitness of each strain, the population size and the number of segregating mutations and substitutions are saved.
Simulation output was analysed using R. For each population we calculated its average fitness, its fitness variance, its dN/dS to the ancestor, using the number of mutations per strains as the differences with the ancestor and the within population dN/dS, using the difference between pair of strains of the same population. The SLiM code is reproduced in supplementary material.

Acknowledgements:
The authors thank Sabrina Woltemate for technical assistance. This work was supported by Sequencing Grants-in-aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan (221S0002 and 18KK0266) to YY. FFV is financed by FCT through Assistant Researcher grant CEECIND/03023/2017. KT and the sequencing of KI isolates was supported by Erik Philip-Sörensen Foundation grant G2016-08, and Swedish Society for Medical research (SSMF). All primary bioinformatics and parts of the comparative genomics were performed on resources provided by Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under projects snic2018-8-24 and uppstore2017270. Work by SS was supported by the German Research Foundation (DFG, project number 158 989 968–SFB 900/A1) and by the Bavarian Ministry of Science and the Arts in the framework of the Bavarian Research Network 'New Strategies Against Multi-Resistant Pathogens by Means of Digital Networking – bayresq.net. DF was supported by Shanghai Municipal Science and Technology Major Project No. 2019SHZDZX02.

References:

1. S. Suerbaum, P. Michetti, Helicobacter pylori infection. *N Engl J Med* **347**, 1175-1186 (2002).
2. S. Schwarz *et al.*, Horizontal versus familial transmission of Helicobacter pylori. *PLoS Pathog* **4**, e1000180 (2008).
3. D. Falush *et al.*, Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582-1585 (2003).
4. B. Linz *et al.*, An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915-918 (2007).

5. K. Thorell *et al.*, Correction: Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas. *PLoS Genet* **13**, e1006730 (2017).
6. Y. Moodley *et al.*, Age of the association between Helicobacter pylori and man. *PLoS Pathog* **8**, e1002693 (2012).
7. F. Maixner *et al.*, The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**, 162-165 (2016).
8. M. Oleastro, R. Rocha, F. F. Vale, Population genetic structure of Helicobacter pylori strains from Portuguese-speaking countries. *Helicobacter* **22**, (2017).
9. G. Hellenthal *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747-751 (2014).
10. B. M. Henn, L. L. Cavalli-Sforza, M. W. Feldman, The great human expansion. *Proc Natl Acad Sci U S A* **109**, 17758-17764 (2012).
11. M. Lynch, J. Conery, R. Burger, Mutational Meltdowns in Sexual Populations. *Evolution* **49**, 1067-1080 (1995).
12. O. K. Silander, O. Tenaillon, L. Chao, Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biol* **5**, e94 (2007).
13. H. J. Muller, The Relation of Recombination to Mutational Advance. *Mutat Res* **106**, 2-9 (1964).
14. K. E. Lohmueller *et al.*, Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997 (2008).
15. Y. B. Simons, G. Sella, The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Curr Opin Genet Dev* **41**, 150-158 (2016).
16. M. Eppinger *et al.*, Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet* **2**, e120 (2006).
17. B. Lafay, J. C. Atherton, P. M. Sharp, Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. *Microbiology (Reading)* **146 (Pt 4)**, 851-860 (2000).
18. M. A. Antezana, R. R. Hudson, Before crossing over: the advantages of eukaryotic sex in genomes lacking chiasmatic recombination. *Genet Res* **70**, 7-25 (1997).
19. S. Bubendorfer *et al.*, Genome-wide analysis of chromosomal import patterns after natural transformation of Helicobacter pylori. *Nat Commun* **7**, 11995 (2016).
20. G. Morelli *et al.*, Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. *PLoS Genet* **6**, e1001036 (2010).
21. S. S. Duncan *et al.*, Comparative genomic analysis of East Asian and non-Asian Helicobacter pylori strains identifies rapidly evolving genes. *PLoS One* **8**, e55120 (2013).
22. S. Wright, The shifting balance theory and macroevolution. *Annu Rev Genet* **16**, 1-19 (1982).
23. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477 (2012).
24. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
25. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).
26. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).
27. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

28. T. Resende, D. M. Correia, M. Rocha, I. Rocha, Re-annotation of the genome sequence of Helicobacter pylori 26695. *J Integr Bioinform* **10**, 233 (2013).
29. P. Ewels, M. Magnusson, S. Lundin, M. Kaller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016).
30. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
31. K. Yahara *et al.*, Chromosome painting in silico in a bacterial species reveals fine population structure. *Mol Biol Evol* **30**, 1454-1464 (2013).
32. Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000).
33. B. C. Haller, P. W. Messer, Evolutionary Modeling in SLiM 3 for Beginners. *Mol Biol Evol* **36**, 1101-1109 (2019).

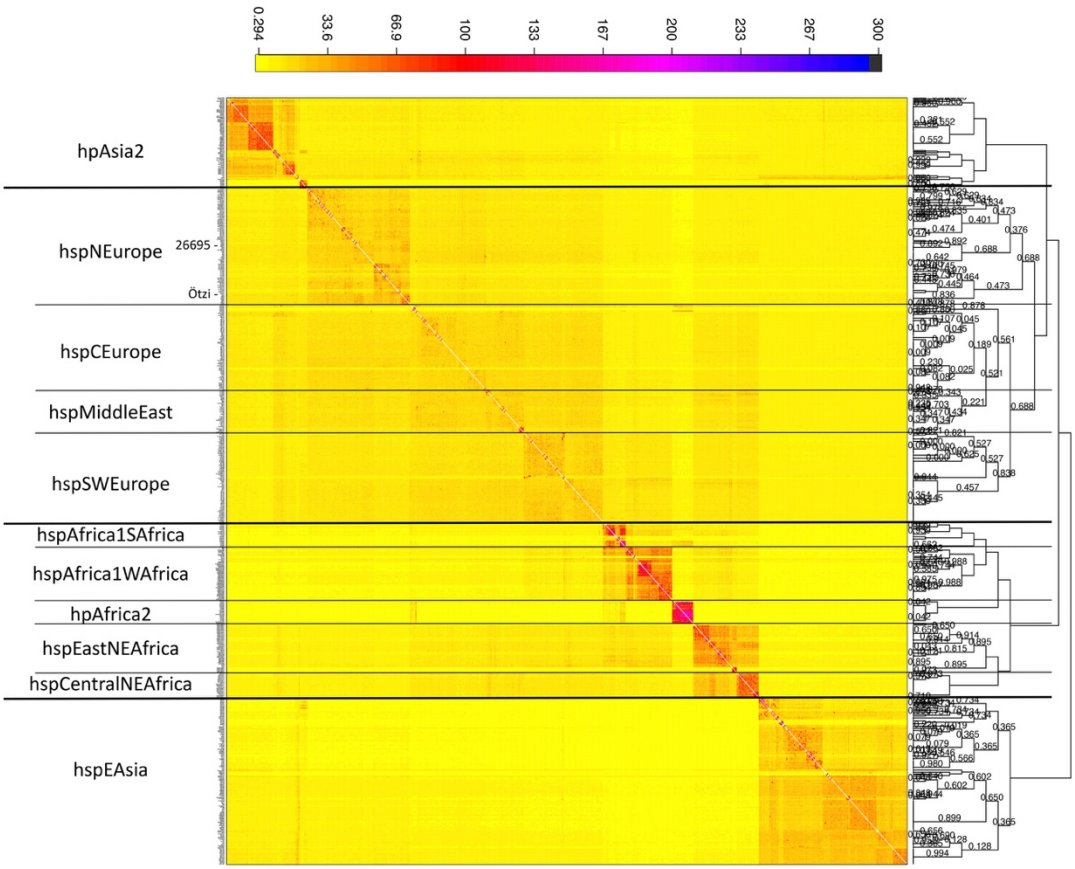

Figure S1 *heatmap showing populations inferred by fineSTRUCTURE. For each individual the number of chunks donated by other individuals are shown according to the colour scale at the top. Note position of Otzi within hpsNEurope. Tree on right shows hierarchical clustering of fineSTRUCTURE populations.*

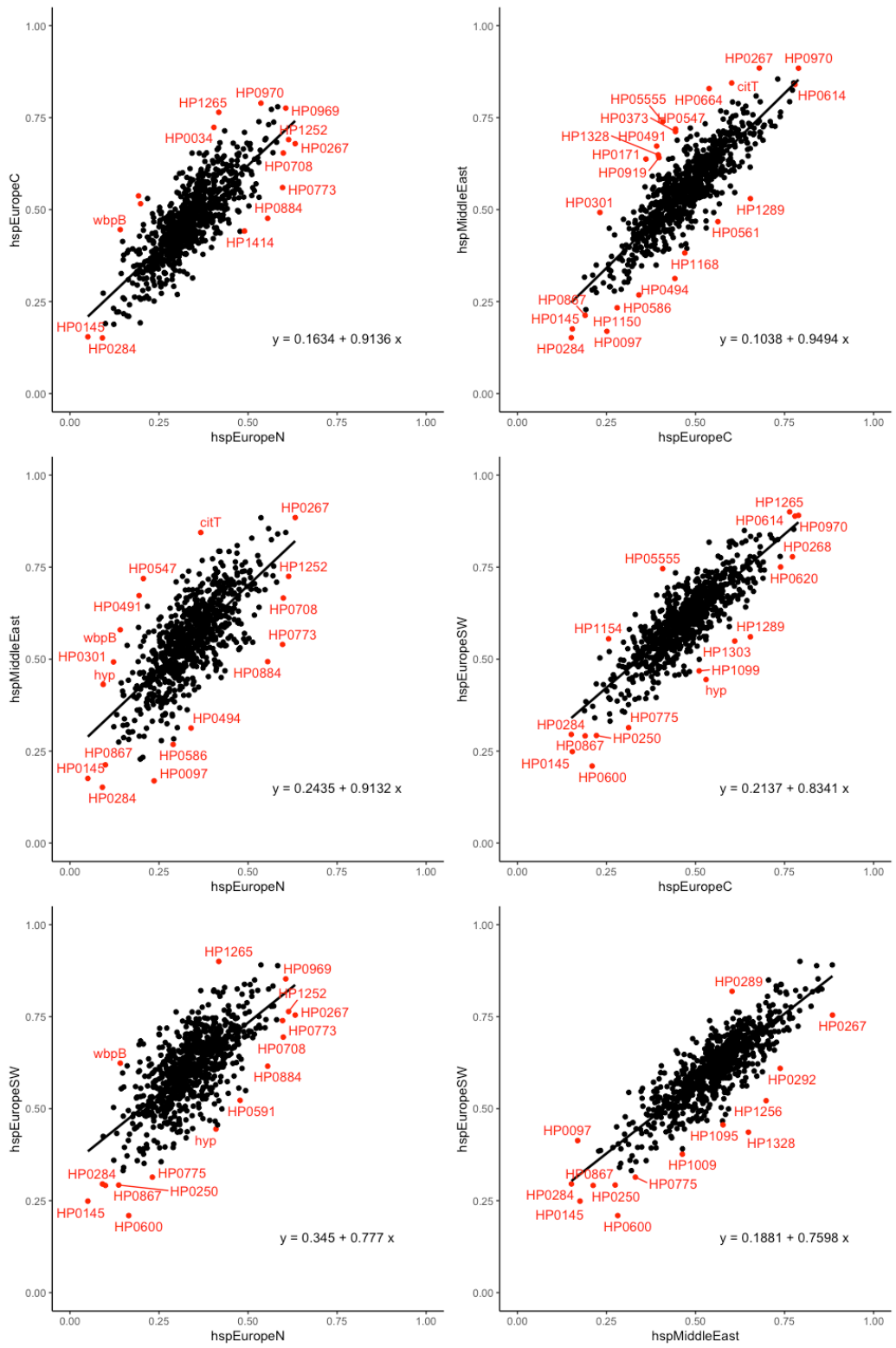

Figure S2 *Average African ancestry proportions of genes.*
Each panel shows relationship between the proportion of ancestry assigned to Africa in the chromosome painting for each gene for pairs of hpEurope subpopulations. Outliers from the regression slope are calculated based on robust Mahalanobis distances using the R function aq.plot(data, quan =1,alpha =0.015).

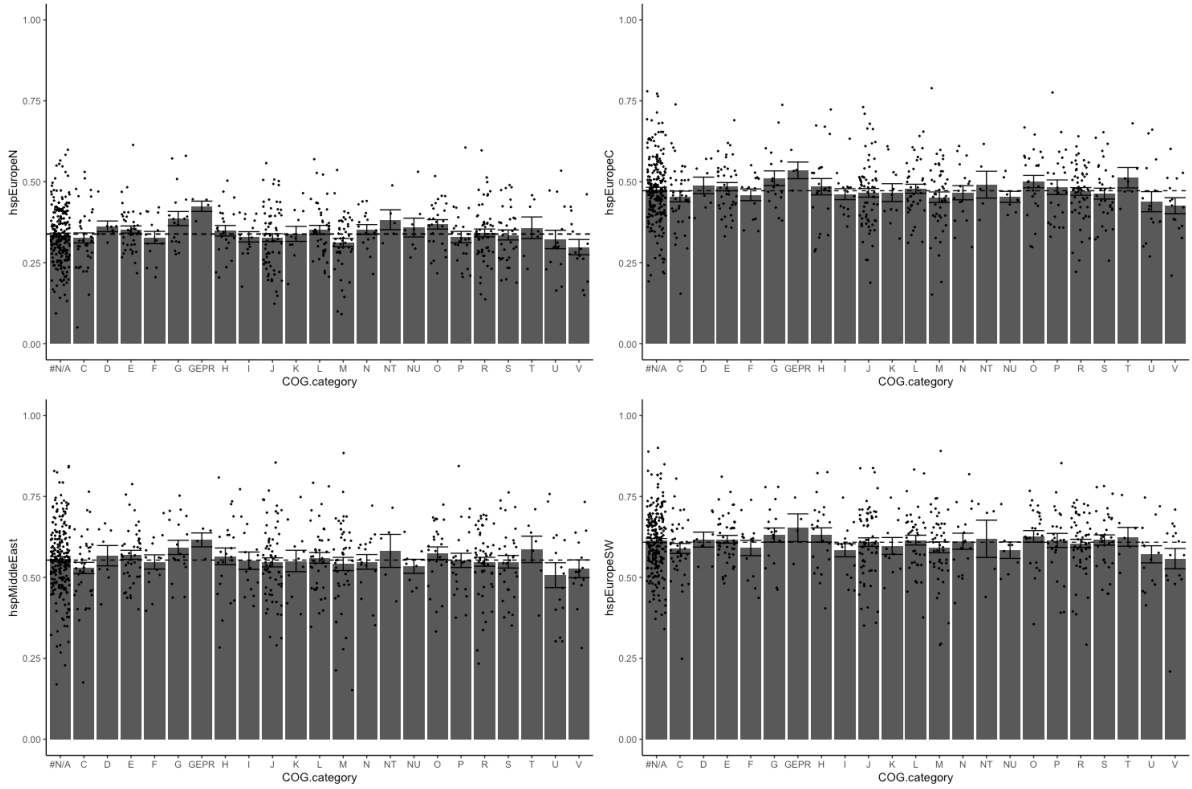

Figure S3 *Average African ancestry proportion of genes in particular COG category. Calculated separately for each hpEurope subpopulation. Whiskers show standard error of the average for each category.*

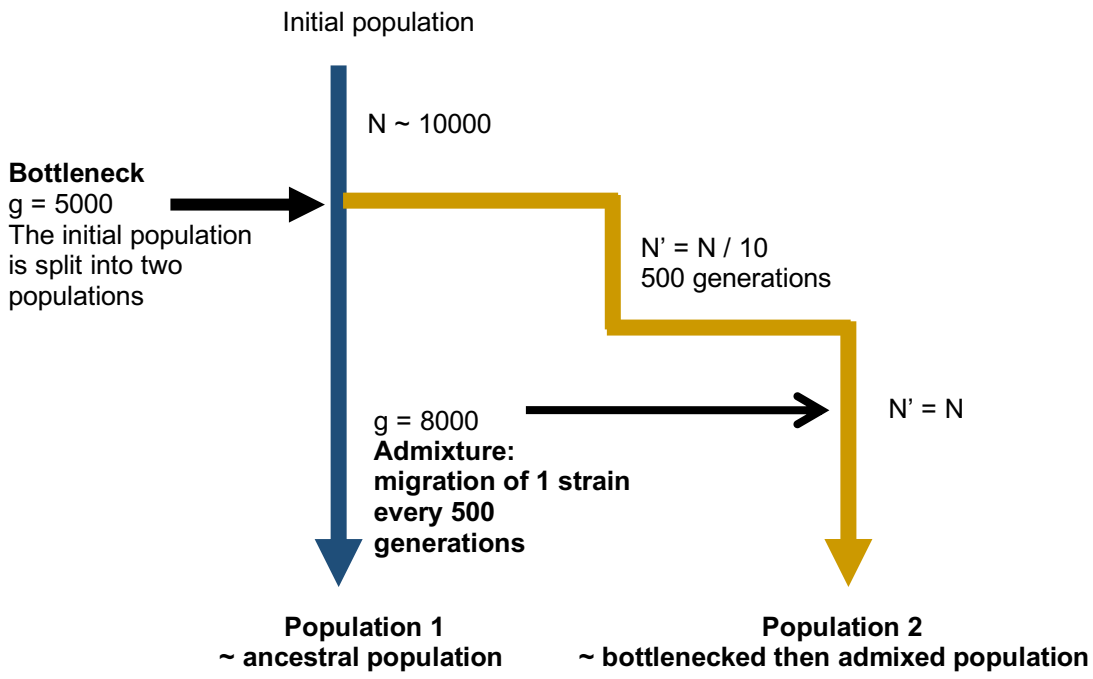

Figure S4 *The different steps of the simulation process.*
Parameters values: N = 10000; genome length = 1.6 Mbp; mutation rate = $5x10^{-7}$ per bp; deleterious mutations: s = (-0.005, -0.002, -0.001, -0.0005, -0.0002, -0.0001); 50% of the mutations (the other 50% are neutral mutations). Look at different recombination levels: clonal reproduction (import size per generation = 0bp), intermediate recombination levels (import size per generation = 500bp and 5000bp) and nearly free recombination (import size per generation = 50000bp).

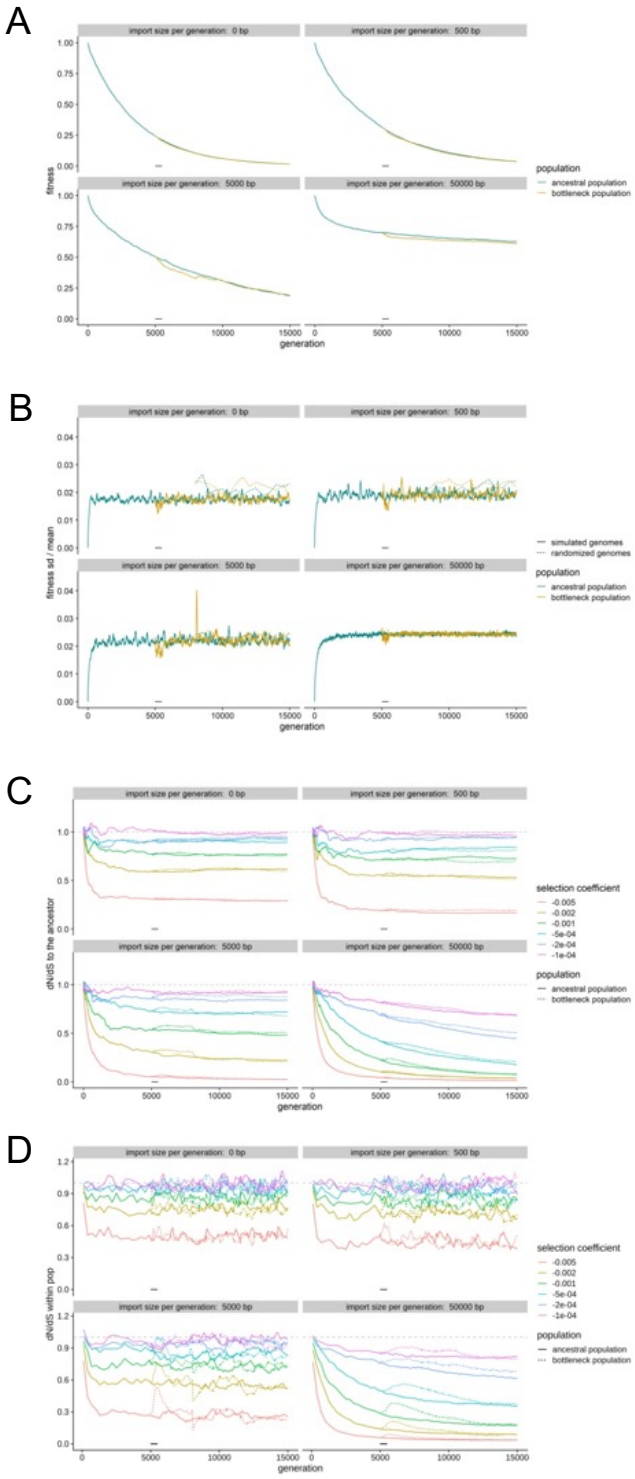

Figure S5 *Simulation of the effects of a bottleneck and admixture for different level of recombination (size of the import for one generation being 0bp, 500bp, 5000bp or 50000bp for a genome of 1.6 Mbp). (A) Average fitness per generation. (B) Fitness standard deviation, normalized by the mean fitness, per generation. (C) dN/dS to the ancestor per generation, for the different classes of mutations. (D) Within population dN/dS per generation, for the different classes of mutations.*

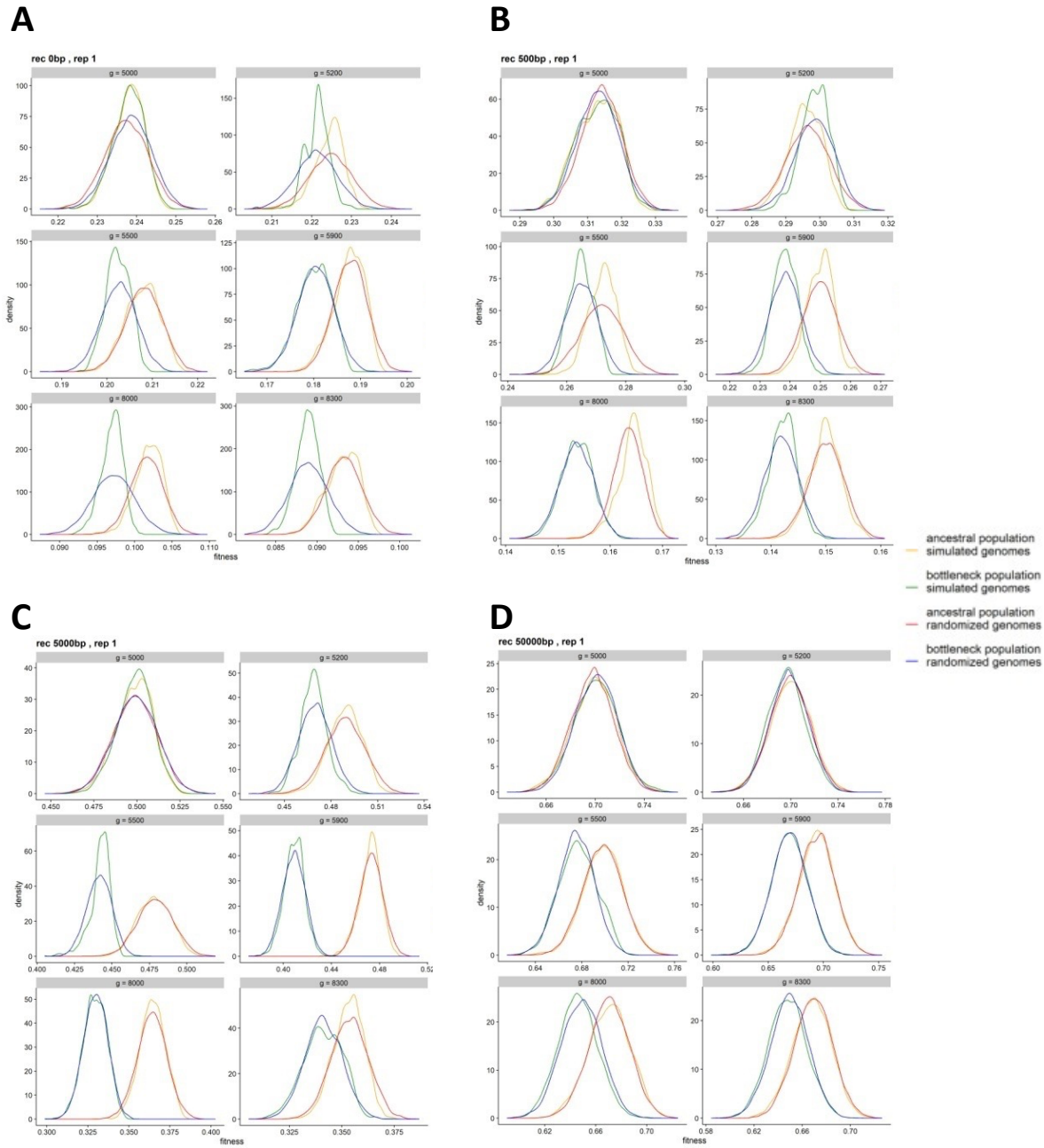

Figure S6 *Differences between the fitness distribution of the simulated populations and a population under free recombination (randomized strains). This comparison is done for both the non-bottleneck population (population 1) and the population subject to the bottleneck and admixture (population 2), for different generations (during and after the bottleneck and for the first admixture event), for an import size per generation of (A) 0bp, (B) 500bp, (C) 5000bp and (D) 50000bp.*

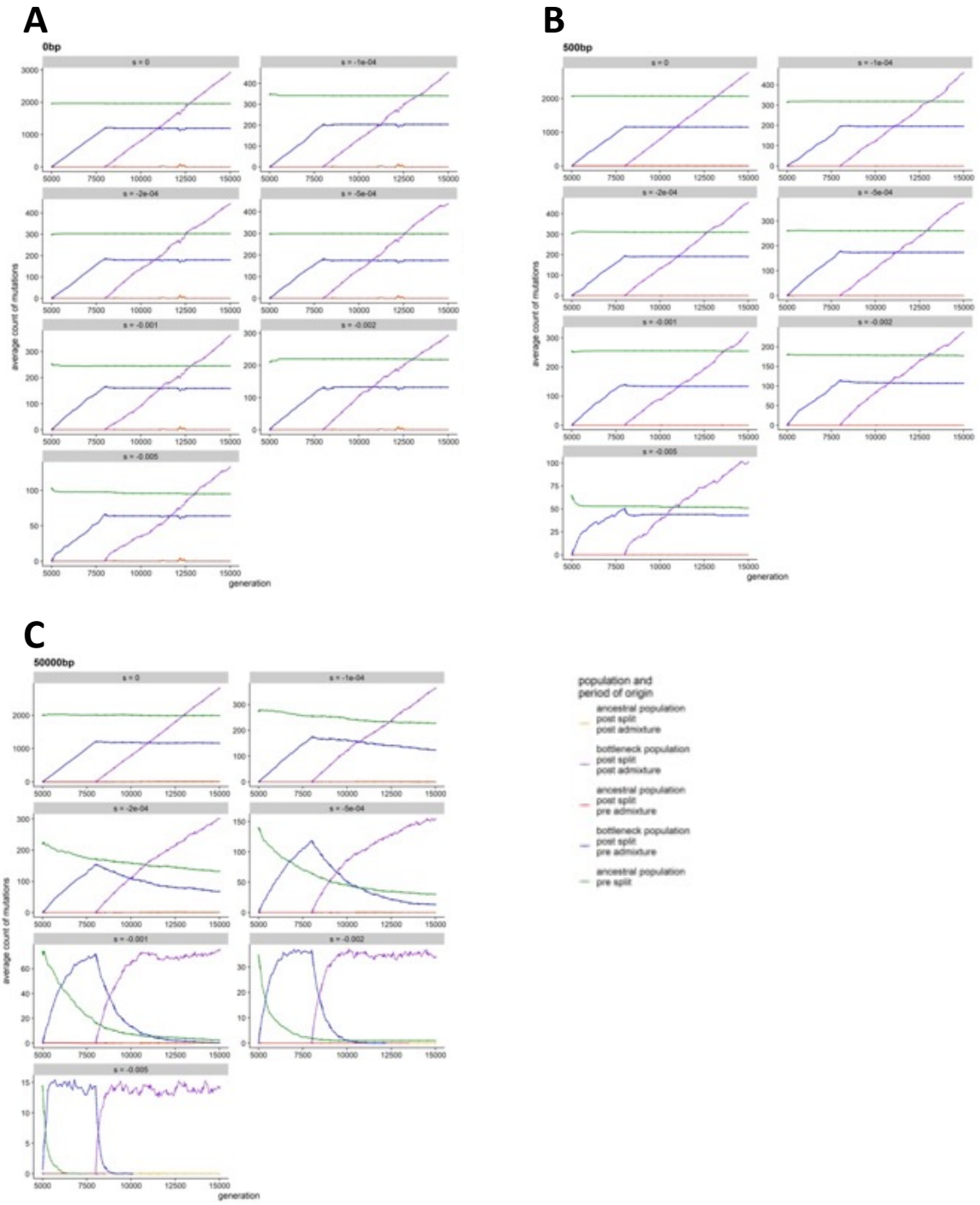

Figure S7 *Average count of mutations in the population subject to bottleneck and admixture, categorized according to selection coefficient and when and where they arose for an import size of (A) 0bp, (B) 500bp and (C) 50000bp.*