Worldwide Population Structure, Long-Term Demography, and Local Adaptation of *Helicobacter pylori*

Valeria Montano,[*,†] Xavier Didelot,[‡] Matthieu Foll,[§,] Bodo Linz,[††,‡‡] Richard Reinhardt,[§§,***] Sebastian Suerbaum,[†††] Yoshan Moodley,[*,‡‡‡,1] and Jeffrey D. Jensen[§,**,1]**

[*]Department of Integrative Biology and Evolution, Konrad Lorenz Institute for Ethology, University of Veterinary Medicine, 1160 Vienna, Austria, [†]Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland, [‡]Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, United Kingdom, [§]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, [**]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [††]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16801, [‡‡]Department of Molecular Biology, Max Planck Institute for Infection Biology, D-10117 Berlin, Germany, [§§]Max Planck Genome Centre Cologne, D-50829 Cologne, Germany, [***]Max Planck Institute for Molecular Genetics, D-14195 Berlin, Germany, [†††]Institute of Medical Microbiology and Hospital Epidemiology, Hannover Medical School, 30625 Hannover, Germany, and [‡‡‡]Department of Zoology, University of Venda, Thohoyandou 0950, South Africa

ABSTRACT *Helicobacter pylori* is an important human pathogen associated with serious gastric diseases. Owing to its medical importance and close relationship with its human host, understanding genomic patterns of global and local adaptation in *H. pylori* may be of particular significance for both clinical and evolutionary studies. Here we present the first such whole genome analysis of 60 globally distributed strains, from which we inferred worldwide population structure and demographic history and shed light on interesting global and local events of positive selection, with particular emphasis on the evolution of San-associated lineages. Our results indicate a more ancient origin for the association of humans and *H. pylori* than previously thought. We identify several important perspectives for future clinical research on candidate selected regions that include both previously characterized genes (*e.g.*, transcription elongation factor *NusA* and tumor necrosis factor alpha-inducing protein *Tipα*) and hitherto unknown functional genes.

KEYWORDS adaptation; neutral evolution; human pathogens

HELICOBACTER pylori is a Gram-negative bacterium that infects the mucosa of the human stomach. It was first described in the 1980s, when it was initially identified in association with chronic gastritis and later causally linked to serious gastric pathologies such as gastric cancer and ulcers (Marshall and Warren 1984; Suerbaum and Michetti 2002). It infects >80% of humans in developing countries and, although its prevalence is lower in developed countries, nearly 50% of the worldwide human population is infected (Ghose *et al.* 2005; Salih 2009; Salama *et al.* 2013).

Due to its clinical and evolutionary importance, there has been considerable research on mechanisms of *H. pylori* transmission, as well as on the population genetics and phylogenetic relationships among global isolates. Thus far, population genetic analyses have mainly focused on seven housekeeping genes (usually referred to as multilocus sequence typing or MLST), with the primary conclusions being that *H. pylori* strains appear highly structured, and their phylogeographic patterns correlate consistently with that of their human hosts. Given that the *H. pylori*–humans association is at least 100,000 years old (Moodley *et al.* 2012), the current population structure of *H. pylori* may be regarded as mirroring past human expansions and migrations (Falush *et al.* 2003; Linz *et al.* 2007; Moodley and Linz 2009; Breurec *et al.* 2011) and thus help us shed light on yet

doi: 10.1534/genetics.115.176404
Manuscript received March 16, 2015; accepted for publication May 15, 2015; published Early Online May 20, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176404/-/DC1.
Data access: NCBI BioProject ID PRJNA245115.

[1]Corresponding authors: EPFL SV IBI-SV UPJENSEN, AAB 0 48 (Bâtiment AAB), Station 15, CH-1015 Lausanne, Switzerland. E-mail: jeffrey.jensen@epfl.ch; and Department of Zoology, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa. E-mail: yoshan.moodley@univen.ac.za

unknown dynamics of local demographic processes in human evolution. However, despite the knowledge gained thus far, the long-term global demographic history of *H. pylori* has never been directly inferred.

The long, intimate association of *H. pylori* with humans suggests a history of bacterial adaptation. Considerable attention has focused on specific genes involved in modulating adaptive immunity of the host (for a review see Yamaoka 2010 and Salama *et al.* 2013) and on genomic changes occurring during acute and chronic *H. pylori* infection (Kennemann *et al.* 2011; Linz *et al.* 2014) as well as during *H. pylori* transmission between human hosts (Linz *et al.* 2013). However, bacterial genome adaptation has not been investigated at the global level. Owing to the recent introduction of next generation sequencing approaches, several complete *H. pylori* genomes have been characterized and are now available to further explore the selective history that might have contributed to shaping the bacterial genome.

Here, we study a combined sample of 60 complete *H. pylori* genome sequences (53 previously published, 7 newly sequenced) with origins spanning all five continents. Our aims were to detect adaptive traits that are commonly shared among the worldwide *H. pylori* population as well as to uncover patterns of local adaptation. We expect that, apart from a generally important role of adaptation to the human gastrointestinal environment, the differing ecophysiological conditions found in the gastric niche of worldwide human hosts, based on diverse diets and different bacterial compositions, could likely generate differential selective pressure on specific bacterial traits leading to locally adaptive events. For instance, an increase in pathogenicity seems to have occurred in *H. pylori* during the colonization of East Asia and could be partially explained by the presence of different alleles of virulence factors (*e.g.*, CagA, VacA, and OipA; Yamaoka 2010); also, colonization of the stomach niche has been optimized by regulation of motility and by bacterial cell shape (Sycuro *et al.* 2012).

To disentangle the signatures of demographic processes from the effects of natural selection on the distribution of allele frequencies, we first investigated the demographic history of our worldwide genome sample. Given that the genetic structure retrieved among the bacterial genomes mirrors the geographic distribution of human populations (Moodley and Linz 2009; Breurec *et al.* 2011; Moodley *et al.* 2012), the vast literature on human demographic history provides a solid basis for the study (*e.g.*, Cavalli-Sforza *et al.* 1994), but modeling human–*H. pylori* coevolution would also require knowledge of transmission dynamics and within-host variation. Despite the large number of surveys carried out, *H. pylori* transmission via an external source has never been demonstrated and direct contact among individuals is still considered the predominant mechanism (Brown 2000; Van Duynhoven and De Jonge 2001; Allaker *et al.* 2002; Perry *et al.* 2006). Transmission also depends on the hosts' access to health care and socioeconomic conditions. In developing countries, *H. pylori* transmission seems to happen preferentially but not exclusively among individuals who are closely related or living together (Schwarz *et al.* 2008; Didelot *et al.* 2013). However, in developed countries, improved hygienic conditions have decreased *H. pylori* prevalence, and transmission occurs primarily between family members, especially from mothers to children (Bures *et al.* 2006; Chen *et al.* 2007; Khalifa *et al.* 2010; Krebes *et al.* 2014). Further, an important epidemiological factor is that a human host is normally infected with *H. pylori* within the first 5 years of life and, unless treated, infection persists the entire host lifespan. The host individual is therefore always potentially infective.

The human stomach is typically infected with a single dominant strain, with multiple infections occurring less frequently (*e.g.*, Schwarz *et al.* 2008; Morelli *et al.* 2010; Nell *et al.* 2013). However, this empirical observation may be due to an experimental approach that intrinsically limits the detection of multiple infections (Didelot *et al.* 2013) since only a single isolate per patient is generally studied, and more focused approaches have highlighted higher within-host variation (Ghose *et al.* 2005; Patra *et al.* 2012). In addition, MLST studies have detected a small fraction of human hosts from the same population sharing the same bacterial strain (or at least highly related strains with identical sequence type) (Patra *et al.* 2012; Nell *et al.* 2013). At the molecular level, mutation and recombination have been identified as the key forces responsible for population genetic variability (Suerbaum and Josenhans 2007). A recent whole genome study on 45 infected South Africans demonstrated that recombination is the major driver of diversification in most (but not all) hosts (Didelot *et al.* 2013), confirming previous observations (Falush *et al.* 2001; Kennemann *et al.* 2011). At the population level, recombination is very frequent throughout the genome along with other events such as rearrangements, transpositions, insertions, and gene gain or loss (Gressmann *et al.* 2005; Kawai *et al.* 2011). The relative roles of demographic and selective processes in shaping the bacterial genetic variation during the lifespan of a single host have yet to be explored.

Given our limited knowledge of *H. pylori* epidemiology and thus its consequences on long-term evolution, we here explore the species' genetic structure using newly available worldwide genomic data to infer the demographic history of the sampled populations, directly addressing the extent to which the population history of *H. pylori* mirrors that of its human host. Using this estimated demographic model as a null, we explore two different approaches to characterize both local and global events of positive selection. Our results indicate global signatures of selection in functionally and medically relevant genes and highlight strong selective pressures differentiating African and non-African populations, with >100 putatively positively selected genes identified.

Materials and Methods

Samples and whole genome sequencing

Seven complete *H. pylori* genomes were newly typed for the present study to increase the currently available set of

53 genomes, to represent all five continents (Supporting Information, Table S1). The most valuable contributions among our sequences were the Australian aboriginal, Papua New Guinean Highlander, Sudanese Nilo-Saharan, and South African San genomes, which have never been previously characterized.

Data production was performed on a Roche 454 FLX titanium sequencer. While genome sequencing (WGS) libraries for pyrosequencing were constructed according to the manufacturer's protocols (Roche 2009 version). Single-end reads from 454 libraries were filtered for duplicates (gsMapper v2.3, Roche) and could be directly converted to frg format that was used in the genome assembler by Celera Assembler v6.1 (CA6.1; Miller *et al.* 2008). Several software solutions for WGS assembly were tested during the project, among them Roche´s Newbler and CeleraAssembler (both can assemble all read types). Genome assembly was performed on a Linux server with several TB disk space, 48 CPU cores, and 512 GB RAM.

Bioinformatics

After long read assembly, the seven new genomes were further reordered using the algorithm for moving contigs implemented in Mauve 2.3.1 software for bacterial genome alignment (Darling *et al.* 2004, 2010). In this analysis, the scaffold sequence for each genome to be reconstructed was assigned on the basis of geographical proximity. In particular, the sequences from Papua New Guinea and Australia were reordered against an Indian reference (*H. pylori* India7, GenBank reference: CP002331.1; see Table S1), given the absence of closer individuals. The global alignment of the genomes was carried out using mauveAligner in Mauve 2.3.1 with seed size calibrated to ~12 for our dataset (average size ~1.62 Mb). The minimum weight for local collinear blocks, deduced after trial runs performed using default parameter settings, was set to 100. The original Mauve alignment algorithm was preferred to the alternative progressive approach (progressiveMauve; Darling *et al.* 2010) because of its higher performance among closely related bacterial genomes (appropriate in the present case of intraspecific analysis), its higher computational speed, and to avoid the circularity of estimating a guide phylogenetic tree to infer the alignment. The aligned sequences shared by all genomes were uploaded into R using the package *ape* (Paradis *et al.* 2004) and processed for postalignment refinement. The length of the genomes prior to alignment ranged from 1,510,564 bp to 1,709,911 bp with an average of 1,623,888 bp. The Mauve alignment consisted of 71 blocks commonly shared by all the individuals for a total of 2,586,916 sites. Loci with >5% missing data were removed, giving a final alignment of length to 1,271,723 sites. The final number of segregating sites in the global sample was 342,574 (26.9%). Among these, we found 302,278 biallelic sites and 35,003 and 5,293 tri- and tetraallelic sites, respectively. The distribution of segregating sites along the aligned sequences is shown in Figure S1.

Structure analysis

To first define the populations to be used in subsequent analyses, we compared a multivariate approach, discriminant analysis of principal components (DAPC) (Jombart *et al.* 2010) with two Bayesian analyses of population structure BAPSv5.4 (Corander *et al.* 2006, 2008) and STRUCTURE (Pritchard *et al.* 2000). The first method assesses the best number of clusters optimizing the between- and within-group variance of allele frequencies and does not assume an explicit biological model, while the second is based on a biological model that can also detect admixture among individuals. The optimal number of population clusters was established by both methods. In DAPC this is done through the Bayesian information criterion (BIC) using the *find.clusters* function in *adegenet* 3.1.9 (Jombart 2008; Jombart and Ahmed 2011), while BAPS estimates the best K comparing the likelihood of each given structure. We ran the DAPC analysis with 1,000 starting points and 1,000,000 iterations and found that results were consistently convergent over 10 independent trials. BAPS was run with a subset of 100,000 SNPs using the admixture model for haploid individuals and was shown to be effective to detect bacterial populations and gene flow in large-scale datasets (Tang *et al.* 2009; Willems *et al.* 2012). STRUCTURE was run on a subset of 100 kb, for a total of 29,242 SNPs, using 10,000 burn-in and 50,000 iterations, and we replicated 5 runs for each tested number of partitions (from 2 to 10) with the admixture model. Finally, the seven housekeeping genes historically used in *H. pylori* population genetics (MLST) were extracted from the alignment and used to assign populations to strains with STRUCTUREv2.3.4 (Falush *et al.* 2003) as a comparison with previous work.

For further insight into population structure, we reconstructed the clonal genealogy of bacterial genomes using ClonalFrame v1.2 (Didelot and Falush 2007). This method reconstructs the most likely clonal genealogy among the sequences under a coalescent model with mutation and recombination, so that the model of molecular evolution takes into account both the effect of mutated sites and imported (recombining) sites. We also evaluated fine-scale population structure from sequence coancestry using fineSTRUCTURE (Lawson *et al.* 2012). This method performs Bayesian clustering on dense sequencing data and produces a matrix of the individual coancestry. Each individual is assumed to "copy" its genetic material from all other individuals in the sample, and the matrix of coancestry represents how much each individual copied from all others.

Population summary statistics (the number of segregating sites, genetic diversity, mean number of pairwise differences, Tajima's D, and pairwise F_{ST}) were estimated with R packages adegenet and pegas (Paradis 2010).

Inferring demographic history

The genomic landscape is shaped by the combined evolutionary signature of population demography and selection.

Not accounting for population demography, therefore, could lead to biased estimates of both the frequency and strength of genomic selection (*e.g.*, Thornton and Jensen 2007). While many of the available statistical methods for detecting patterns of genome-wide selection have been argued to be robust to demographic models of population divergence and expansion (Nielsen *et al.* 2005; Jensen *et al.* 2007b; Foll and Gaggiotti 2008; Narum and Hess 2011), they also have limitations (Narum and Hess 2011; Crisci *et al.* 2013). In highly recombining species such as *H. pylori* (Morelli *et al.* 2010; Didelot *et al.* 2013), evidence of recent positive selection events across the global population may have become obscured, owing to the reduced footprint of selection.

It was therefore necessary to first explicitly infer the demographic history, to disentangle the effects of population demography on the allele frequency distribution from the possible effects of selective processes. Here, we tested different neutral demographic scenarios, making assumptions based on the observed genetic structure and previous knowledge of human evolutionary history.

Demographic scenarios were modeled and implemented in the software *fastsimcoal2.1* (Excoffier *et al.* 2013), allowing for the estimation of demographic parameters based on the joint site frequency spectrum of multiple populations. The software calculates the maximum likelihood of a set of demographic parameters given the probability of observing a certain site frequency spectrum derived under a specified demographic model. This program uses nonbinding initial search ranges that allow the most likely parameter estimates to grow up to 30%, even outside the given initial search range, after each cycle. This feature reduces the dependence of the best parameter estimates on the assumed initial parameter ranges. Model details and initial parameter range distributions are given in File S1 and File S2. We assumed a finite site mutation model, meaning that the observed and simulated joint site frequency spectra were calculated to include all derived alleles in multiple hit loci (Figure S6).

Model choice and demographic estimates

First, different tree topologies based on hierarchical structures, as obtained with the approaches described above, were compared to infer the best population tree, assuming divergence without migration. Once the tree topology with the strongest statistical support was established, we evaluated and compared the likelihood of models including asymmetric migration among populations. Migration models were tested starting with interchanging individuals only among single pairs of closely related populations. We could therefore assess whether adding migration would improve the likelihood compared to a divergence-without-migration model, and which pairs of populations are most likely to exchange migrants. We also allowed migration among more distantly related populations in addition to a simple pairwise stepping stone model.

The best model among those tested was selected through the corrected Akaike Information criterion (AICc) based on the maximum likelihoods calculated for independent runs.

Testing models of positive selection

Two different statistical tests were used to detect global and local candidate loci for selection. First we used the SweeD algorithm (Pavlidis *et al.* 2013), derived from SweepFinder (Nielsen *et al.* 2005) to localize recent events of positive selection, an approach based upon comparison with the "background" site frequency spectrum (SFS) (Figure S7). The scan for positive selection is carried out by centering the maximized probability of a selective sweep on a sliding-window locus along the chromosome, and calculating the composite likelihood for each centered locus to fall within a region where the distribution of SNPs deviates from the neutral expectation. When an outgroup sequence is available to establish derived mutations, the empirical site frequency spectrum estimated from the observed dataset is unfolded, otherwise only minor alleles are used for the calculation (*i.e.*, a folded SFS). Given the difficulties associated with bacterial genome alignment of suitably close outgroup species, we ran our estimates on a folded SFS. All tri- and tetraallelic SNPs were removed, and monomorphic loci were not considered in the calculation and the grid was set to 500,000 bp. We analyzed the entire dataset (60 genomes) as well as each of the five populations separately.

Second, we applied a method based on the detection of patterns of linkage disequilibrium (LD) around a SNP (OmegaPlus) (Kim and Nielsen 2004; Jensen *et al.* 2007a; Alachiotis *et al.* 2012), since LD is expected to result from a selective sweep owing to the hitchhiking of linked neutral mutations (Maynard Smith and Haigh 1974). This complements the SFS approach as it is applicable to subgenomic regions, contrary to SweeD, and it has proven effective under specific demographic models for which SFS-based approaches are less powerful (Jensen *et al.* 2007a; Crisci *et al.* 2013). We used windows of size between 1,000 and 100,000 bp.

Finally, a total of 1,000 simulated datasets, generated using most likely demographic parameter estimates, were analyzed with SweeD and OmegaPlus to gain an empirical distribution of likelihoods (SweeD) and omega values (OmegaPlus) in a neutrally evolving population. The only parameter drawn from a range was the recombination rate, calibrated around the most likely estimate obtained with ClonalFrame, with the aim of providing an empirical evaluation of its impact on the methods we used to infer selection. The simulated distribution of these selection statistics, based upon the previously inferred demographic history, allows for statistical statements to be made regarding the likelihood that observed outliers are consistent with neutrality alone. A *P*-value for each observed omega and likelihood was obtained using the function *as.randtest* of *ade4* R package, calculated as (number of simulated values equal to or greater than the observed one + 1)/(number of simulated values + 1).

Gene annotation and biological interpretation of the results

Annotation of the bacterial genes was performed using the free automated web server *BASys* (Bacterial Annotation System, www.basys.ca) (Van Domselaar *et al.* 2005). The annotation was run on aligned sequences, removing multiply hit loci. The annotated genome of Africa1 is provided as an example in File S3, and all annotation files are available upon request. The regions identified as being under selection were then compared with the gene annotation.

Results

Population structure and genetic diversity

Given the difficulties of defining a population among a bacterial sample, we decided to perform our cluster analysis using three approaches (DAPC, BAPS, and STRUCTURE) that rely on very different assumptions, keeping in mind that using semi or fully parametric methods (such as STRUCTURE-like approaches) is more likely to lead to violation of the methodological assumptions and therefore to biased results (Lawson 2013). DAPC may outperform STRUCTURE when dealing with datasets with a high degree of isolation by distance (*e.g.*, Kalinowsky 2011), as it is likely the case for *H. pylori* populations (Linz *et al.* 2007; Moodley and Linz 2009), and it also provides the possibility of visualizing clusters' reciprocal distances in the multivariate discriminant space. BAPS and STRUCTURE, on the other hand, offer a biological model to test individual admixture, which is particularly useful to gain an understanding of the degree of differentiation, such that these methodologies may be considered complementary. Population structure analyses were consistent between the model-free DAPC and model-based BAPS and STRUCTURE approaches. All structure approaches were in agreement on a worldwide number of populations that does not exceed $K = 4$. DAPC indicated $K = 4$ as the best clustering (Figure S2A) while BAPS estimates $K = 3$ and STRUCTURE analysis offers a best K in between 2 and 4, with most support for $K = 3$ and partitions >5 dramatically decreasing the likelihood (Figure S2B). Most importantly, the three methods are in consistent agreement on the assignment of single individuals to clusters (Table S1). With the least hierarchical division ($K = 3$), one population comprised African genomes containing all strains from Khoisan-speaking human hosts (referred to as Africa2; Figure 1A and Figure S3). Other African and European strains fell into a population cluster, called here AfricaEu (Figure 1A and Figure S3). A final population is composed of Central Asian, Sahul, East Asian, and Amerind strains (AsiaAmerica; Figure 1A and Figure S3). Finer structuring ($K = 4$) separates the non-Khoisan African sequences (Africa2 and Africa1), but merged European with Central Asian sequences into a new population (referred to as EuroAsia), with Asian and American strains making up the fourth cluster (AsiaAmeria). The only difference between DAPC, BAPS, and STRUCTURE analyses at $K = 4$ is given by individual 7, which is clustered in the AsiaAmerican or EuroAsian populations, respectively. At $K > 4$, American strains were separated into a fifth independent cluster by DAPC, but not by BAPS or STRUCTURE. Plotting the first two discriminant components (DCs) for $K = 4$ (Figure 1B) most strikingly depicted the second African cluster as highly divergent along DC1, whereas divergence among the other clusters was mainly along DC2.

The clonal genealogy (Figure S4) and analysis of fine structure (Figure S5) were in strong agreement with the geographical structuring elucidated by previous approaches. The Africa2 population was well differentiated in the genealogical tree (Figure S4) and in the coancestry matrix (Figure S5), while the remaining populations appear more closely related, and all non-African strains formed a clearly monophyletic clonal group. Asian and American populations were well differentiated in the coancestry analysis and were divided into distinct subclades in the clonal genealogy. The two Sahul genomes shared a higher degree of relatedness with three Indian genomes and these did not cluster monophyletically with the other Eurasian genomes in the clonal genealogy, instead clustering geographically between Eurasian and East Asian groups (see both Figure S4 and Figure S5). Individual 7 appeared intermediately related to both Indian-Sahul and the more divergent Amerind strains. In the following analyses, this strain was left within the European population as indicated by BAPS and also by STRUCTURE analyses of the MLST data (hpEurope).

The population genomic structure elucidated here is in agreement with previous analyses of global structuring of MLST genes, where the highest diversity was found among African strains, the most divergent being the population hpAfrica2 (Falush *et al.* 2003). They also agree that Central Asian (hpAsia2), North-East African (hpNEAfrica), and European (hpEurope) strains are closely related (Linz *et al.* 2007) and sister to hpSahul (Australians and New Guineans) (Moodley *et al.* 2009), and that East Asian and Amerind strains (hpEastAsia) share a relatively recent common ancestor (Moodley and Linz 2009). The divergent hpAfrica2 was shown to have originated in the San, a group of click-speaking hunter-gatherers whose extant distribution is restricted to southern Africa (Moodley *et al.* 2012). A complete list of individuals, geographic origin, and cluster assignment based on DAPC, BAPS, and STRUCTURE (100 kb and MLST extracted from our alignment) is given in Table S1. Predictably, genetic diversity indices were highest for the Eurasian population containing the geographically diverse strains from North East Africa-Europe-Central Asia and Sahul, especially evident from the number of triallelic and tetraallelic loci and the mean number of pairwise

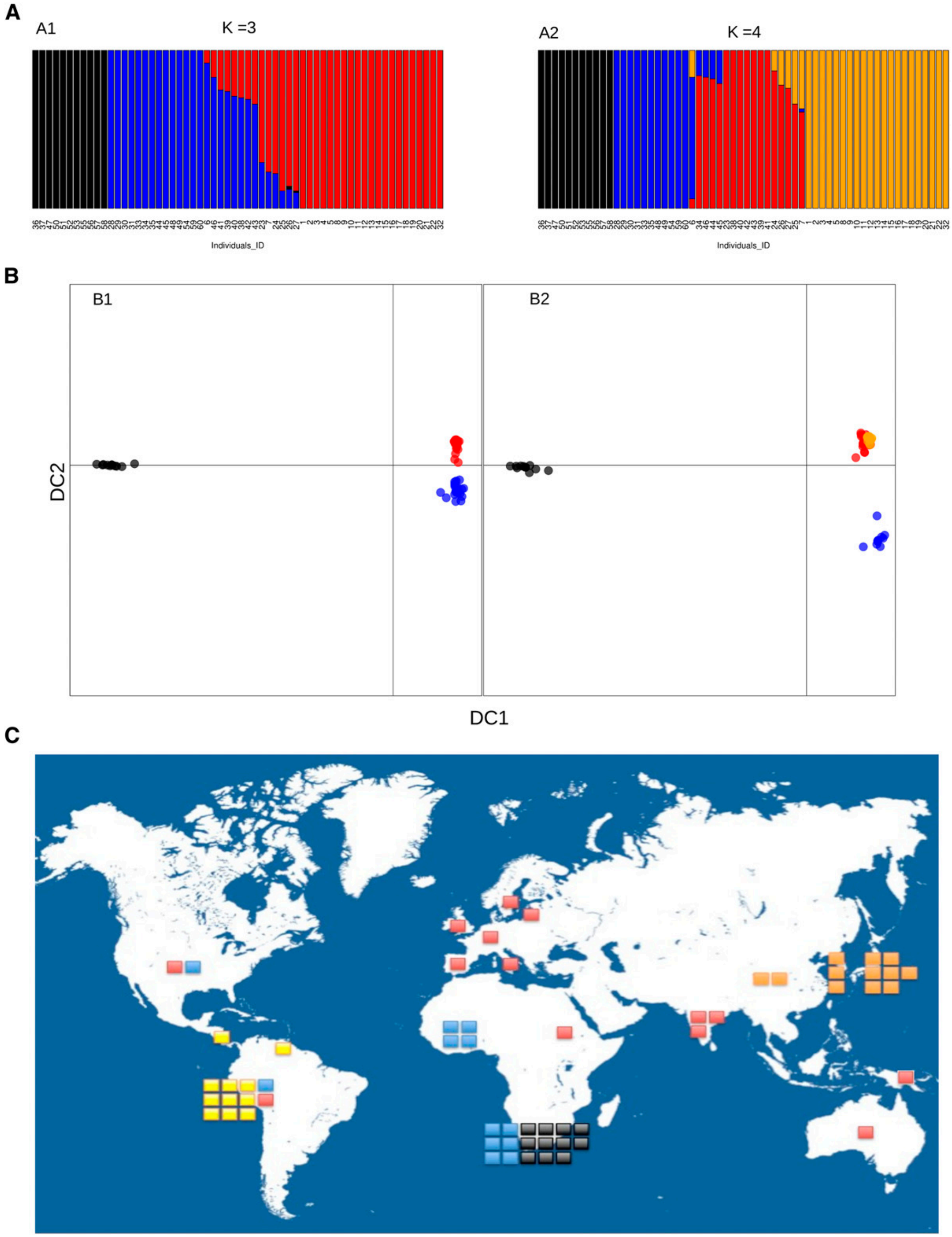


Figure 1 (A) Plots of individual assignments to clusters according to BAPS and DAPC analysis, using *K* = 3 (A1) clusters: black, Africa2; blue, AfricaEu; red, EuroAsia and *K* = 4 (A2) clusters: black, Africa2; blue, Africa1; red, EuroAsia; and orange, AsiaAmerica. (B) Scatterplots of the discriminant space (components 1 and 2), using *K* = 4 (B1), *K* = 5 (B2). (C) World map with squares representing individuals colored according to cluster assignments with yellow squares indicating American subcluster (as for *K* = 5 in DAPC analysis; see Table S1).

differences, while the Amerind population was most homogeneous (Table 1). It is worth noting that within the EuroAsian population there is the highest nucleotide diversity, as European sequences show a value of 0.042 (±0.0005), the three Indian strains 0.038 (±0.0008), and the only two Sahul sequences 0.036 (±0.0013). Only

Table 1 Population summary statistics based on a globally representative dataset of 60 *Helicobacter pylori* genomes

		Number of segregating loci					
Population	N	2 alleles	3 alleles	4 alleles	n	Tajima's D	P-value
Africa2	11	117,125	4,276	171	40472.9	−0.40437	0.747
Africa1	12	139,958	7,034	345	50885.2	−0.03660	0.995
EurAsia	16	197,093	16,713	1,325	63187.7	−0.52261	0.649
EastAsia	12	127,160	5,508	275	40246.74	−0.79713	0.476
America	9	101,895	3,777	183	30781.8	−0.47308	0.706

N is the number of strains per population; *n* is the mean number of pairwise differences. *P*-values refer to Tajima's *D*.

the Africa1 population reaches such value of internal diversity (0.038 ± 0.0003), while all the others fell below 0.03.

Demographic inference

Overall, the different clustering methods and genealogical approaches implemented here were largely consistent in their population assignment. Although the American cluster appears to be more likely substructure, we included it into the further analyses as a separated population. This is owing to the fact that the demographic and selective history associated with the peopling of the Americas would suggest that this group of strains have likely undergone a very different fate than the East Asian strains with which they are closely related. This notion seems indeed to be confirmed by the population-specific tests of positive selection presented below. Furthermore, treating American strains separately offers the possibility of testing the hypothesis of a concerted bacterial–human expansion, as the timing of human colonization of the Americas is a well-characterized event, allowing for comparison with our inference. We proceeded hierarchically to test different genealogical topologies building on the population structure outlined above. First we tested the hypothesis of three main worldwide populations ($K = 3$, panel A, Figure S5), with Africa2 strains forming the most ancestral population, in agreement with our and previous findings (Moodley *et al.* 2012). Alternative origins of the two other clusters—AfricaEu and AsiaAmerica—were therefore tested in three possible topologies (1–3, panel A, Figure S5), with these two populations derived after an ancient split with the Africa2 ancestral population (Figure S6). A comparison of likelihoods suggests the first genealogical setting (see Figure S6) as the most supported, that is, AfricaEu strains are more ancestral than Eastern Asian and American strains, following a pattern close to that of human expansion (Table S2A).

Introducing a further population subdivision (*i.e.*, $K = 4$), we tested different hypotheses for the origin and timing of the out-of-Africa subpopulations, that is EuroAsia and AsiaAmerica (Figure S6B). Lastly, we considered an additional subpopulation formed by American strains, in agreement with DAPC subdivision at $K > 4$ (Figure S6C). Clearly, the addition of multiple populations decreases the degrees of freedom and likelihood value of demographic models, and the hierarchical levels A, B, and C are thus not directly comparable. However, in all tests, a model of population split resembling human expansion out of Africa was always preferred (Table S2A). The results of demographic inference for models without migration were highly compatible across different population substructures (Table S2A).

Finally, hierarchical models based on five populations, and using the most likely genealogical topology obtained with a purely divergent model, were also tested under the assumption of asymmetrical between-lineage gene flow. Each time a pairwise asymmetric migration rate improved the likelihood of the model, the same scenario was reanalyzed adding a further pairwise migration rate, for a total of 20 demographic models tested (divergence plus migration). Pairwise migration rates among populations improved the likelihood of the divergence model, and the addition of further interpopulation migrations highlighted that the most likely model is an asymmetric full island, although this model supports very little gene flow among these major worldwide populations (consistently $<<0.001$ of effective population size per generation; Table 3). The corrected AIC takes into account both number of parameters and number of observations, allowing for a consideration of differences in the likelihood comparison (Table S2). We ran these demographic inferences with and without redundant (near identical) genome sequences from populations Africa2 and Africa1 (30, 31, 48, and 53) to correct for potential sampling bias, and obtained highly similar results.

Comparing population parameters estimated with different models indicates that the introduction of migration primarily influences results concerning the time of population splits and mutation rate (Table 2). While effective past and current population sizes have different absolute values, trends of population reduction (African populations) and growth (non-African populations) are confirmed throughout different models.

The timing of the two population splits, T2 and T4 (Figure 2, Table 2A), which presumably correspond to the out-of-Africa and American colonization events, are comparable to human estimates of population splits. Indeed, the second event appears to be two to four times more recent than the first (on average, ~38,000 generations *vs.* ~110,000 generations, respectively), as expected under a bacterial–host model of coexpansion. According to models without migration, the estimate of divergence in number of generations of the Africa2 population from the other African strains (T1) also fits the timing of the divergence of the San

Table 2 Most likely demographic parameters estimated with fastsimcoal2.1 for tree topology 2 and relative confidence intervals calculated for the migration model, which is the most supported

A. Ancestral population sizes from population Africa2 (Na0) to America (Na4) and current population sizes from population Africa2 (Nc0) to America (Nc4)

Parameters		Ancestral effective populations size (N_a)					Current effective population size (N_c)				
Model	Populations	Na0	Na1	Na2	Na3	Na4	Nc0	Nc1	Nc2	Nc3	Nc4
No migration	Three	1,918,265	309,326	171,636	—	—	106,580	898,799	3,178,763	—	—
No migration	Four	1,852,974	809,566	477,655	51,188	—	101,030	230,367	2,307,144	1,875,418	—
No migration	Five	2,373,071	1,083,279	293,795	28,444	18,449	103,453	257,756	2,015,267	761,583	690,042
Migration	Five	1,215,564	223,068	10,366	11,580	11,821	102,164	106,142	813,228	399,904	184,468
c.i. 0.05	Five	240,056	12,066.4	22,252.0	12,017	11,990	105,459	103,772	230,774	108,383	128,091
c.i. 0.95	Five	2,056,863	395,493	452,400	357,243	383,897	1,154,059	841,312	1,375,820	817,227	789,851

B. Time since population split from most ancient (T1) to most recent (T4), growth rates from population Africa2 (r0) to America (r4) and mutation rate

Population splitting times				Population growth rates					Mutation rate
T1	T2	T3	T4	r0	r1	r2	r3	r4	μ
273,339	138,190	—	—	1.057e-05	−3.902e-06	−2.112e-05	—	—	8.069e-06
229,697	119,441	75,369	—	1.267e-05	5.472e-06	−1.318e-05	−4.778e-05	—	4.596e-06
245,942	128,714	45,670	31,778	1.274e-05	5.837e-06	−1.496e-05	−7.198e-05	−0.000113	1.479e-07
529,626	89,686	69,096	44,338	4.675e-06	1.402e-06	−4.864e-05	−5.126e-05	−6.196e-05	0.0009732
102,889	53,942	34,197	11,430	−6.223e-06	−9.437e-06	−3.978e-06	−7.891e-05	−1.693e-04	0.0002284
350,810	95,913	51,312	28,388	1.434e-05	5.251e-06	−3.558e-05	7.689e-06	1.172e-05	0.0008439

Parameters are reported assuming two generations per year. Population parameters correspond to those depicted in Figure 3; *r* parameters are the population growth rates, with the numeric order indicating populations from Africa2 to America (see Figure 3); μ is the mutation rate.

population from other Africans, being twice as old as the out-of-Africa divergence (~249,000 generations ago) (Table 2A). Indeed, previous inferences based on human genetic data have estimated these events to have happened ~60 kya for the out-of-Africa (Eriksson *et al.* 2012), ~20 kya for the arrival into the Americas (Eriksson *et al.* 2012), and ~110 kya for San divergence (Hammer *et al.* 2011; Veeramah *et al.* 2011; Schlebusch *et al.* 2012). On the other hand, the time inferred from the *H. pylori* dataset for the San split under the most likely model, which includes migration, is older than ~500,000 generations.

The long-term mutation rate per site per generation estimated with *fastsimcoal2.1* varies between ~8.47 × 10^{-7} and ~9.73 × 10^{-4} (Table 2), this second estimate being much faster than the previous long-term estimate, per site per year, from Morelli *et al.* (2010), based on the coalescent tree of the seven housekeeping genes and inferred with ClonalFrame (2.6 × 10^{-7}). Other previous estimates based on 78 gene fragments from serial and family isolates (1.4–4.5 × 10^{-6}) (Morelli *et al.* 2010), upon genomes sequentially taken from patients with chronic infection (2.5 × 10^{-5}) (Kennemann *et al.* 2011) and on genomes from 40 family members (1.38 × 10^{-5}) (Didelot *et al.* 2013) are compatible with that inferred here by a purely divergent model. The bacterial recombination rate per initiation site per year obtained from our genomes analyzed with ClonalFrame (9.09 × 10^{-9}) is >20 times slower than a previous estimate of 2.4 × 10^{-7} reported in Morelli *et al.* (2010), based on housekeeping genes using the same approach. It is important to note, however, that the recombination rate was not included in our models and that our absolute estimates are in generations instead of years.

Growth rates (*r*, see Table 2A) were negative for African clusters indicating population size reductions, with current effective population sizes (N_c) being several times lower than ancestral population sizes (N_a) for Africa2 and Africa1, respectively (Table 2A). The other three populations show signatures of expansion and appear to have been founded by a comparable few individuals, subsequently undergoing rapid growth. Migration rates are similarly small among pairwise populations, however outgoing migration rates from Africa are lower than the others (Table 3). This result may indicate that gene flow did not extensively involve geographic macroareas, but if it did occur, mixed stains are more likely to be found in specific contact regions (*e.g.*, coastal areas). Confidence intervals of demographic estimates with migration obtained using parametric bootstrap are reported in Table 2 and show important uncertainty associated with the best estimates.

Tests of positive selection and identified candidate regions

After correction of likelihood values with demographic simulations, the SweeD test of selection did not identify any strongly selected loci at the global level (Figure 3), but did indicate differential signatures of positive selection at the population level (Figure 3; Table 4). The largest number of selected loci was detected among African bacterial strains associated with San-speaking people (Africa2). Signatures of local positive selection were also observed in the Africa1

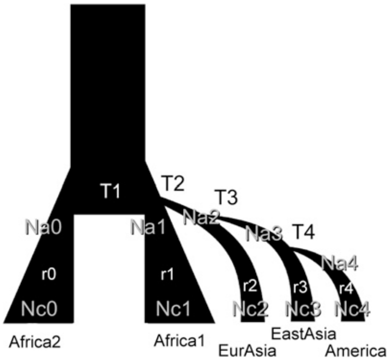


Figure 2 Schematic representation of the most likely genealogy inferred for *H. pylori* worldwide sample. Demographic parameters estimated via coalescent simulations are summarized. *T* parameters correspond to time of population splits (1 to 4, most ancient to most recent). N_a and N_c parameters indicate effective ancestral and current population sizes, with 0 being the Africa2 population and 5 the America population (most ancient to most recent). *R* parameters refer to population growth.

and American populations (Figure 3), while remaining populations (Eurasian and East Asian) did not show strong evidence for recent local adaptation (Figure 3).

The same dataset analyzed with OmegaPlus, using as a null distribution the same demographic simulations analyzed with SweeD, gave different results, with significance found mainly in the worldwide sample (Figure 3). The highest values of linkage disequilibrium were found in the global dataset (Table S3), with the highest peak associated with a gene coding for the elongation protein NusA, which has been studied in *Escherichia coli* (Cohen *et al.* 2010). Despite the structured nature of the worldwide sample, previous studies have demonstrated that population structure has little to no impact on the specific LD structure captured by the Omega statistic (*e.g.*, Jensen *et al.* 2007a).

Both methods, SweeD and OmegaPlus, indicate several signatures of positive selection in African and American populations, while much lower signals are observed for Euro-Asian populations. The synthesis of the two analyses is presented in Figure 3. Regions that were significant for only one of the two methods were considered if their likelihood or omega value overcame the maximum value found for overlapping regions.

Using this approach, 158 genes are identified as putatively positively selected in either the total worldwide datasets or in the five subpopulations (Table S3 and Table S4), with the highest number (51) found in the Africa2 population. Moreover, this includes several unknown genes, most of which appear to code for outer membrane proteins (Table S4). Copper-associated genes (2 *copA* and 1 *copP*) are also indicated as positively selected. These genes are part of the *sro* bacterial operon and may relieve copper toxicity (Table S3; Beier *et al.* 1997; Festa and Thiele 2012). Among Africa2 strains, the highest likelihood values among Africa2 strains correspond to a well-known division protein gene (*ftsA*) (Figure 3 and Table S4). Moreover, the *pyrB* gene coding for aspartate carbamoyltransferase is also identified and was previously suggested as essential for bacterial survival (Burns *et al.* 2000). In the Africa1 population, the most important signal of selection appears associated to a *vacA* gene, a trait that has been consistently studied given its role in the *H. pylori* pathogenic process (*e.g.*, Basso et al. 2008; Yamaoka 2010). Other *vacA* and *vacA*-like genes are indicated in Africa2 and EuroAsian populations (Table S3 and Table S4).

Discussion

Our analysis of a global *H. pylori* genome sample sought to illuminate both the selective and demographic histories of this human pathogen. Our analyses of population structure were carried out with particular attention, as population genetic clusters were the basic unit for demographic and selection inferences. Previous work based on MLST sequences and STRUCTURE software found a higher number of clusters distributed worldwide, a result largely accepted in the field. However, given the importance of population structure and the theoretical and computational limitations of some approaches, as well as the clonal reproductive behavior of our organism, we explored population structure from complementary points of view (*i.e.*, multivariate analysis, Bayesian analysis, coancestry analysis, and coalescent genealogy). This combination of multiple approaches identified fewer populations globally and thus offers an alternative perspective to previous results. Furthermore, our inferred mutation rate represents the first attempt to study the long-term substitution rate of *H. pylori* on a worldwide genome sample. Under a purely divergent model, the result was similar to the long-term rate previously estimated from MLST housekeeping genes (Morelli *et al.* 2010), but introducing migration led to much higher estimates.

While this analysis based on high-resolution data provides a reliable relative estimate of times to population divergence events, the open question remains on how to interpret and compare the bacterial inferences with those based on human genetics. Times of population splits T1, T2, and T4 are, in terms of the number of generations, roughly twice as old as has been proposed in the human demographic literature. If we use these estimates as calibration points to translate number of generations into years, we can deduce a number of bacterial generations per year = 2. An exception is represented by the estimate of San bacterial divergence when migration is accounted for, as the number of generations doubles to ~530,000 translating into ~265 kya of split (still assuming a bacterial number of two generations per year). Notably, one recent estimate of San divergence obtained by Excoffier *et al.* (2013) is very near our estimate, *i.e.*, ~260 kya. If we alternatively used the latter estimate of split of Africa2 strains from others as a calibration point to deduce the number of bacterial generations per year, then we would consider that ~530,000 bacterial generations happened within ~110 kya (which is the most supported estimate of San split from human genetic data). In

Table 3 All *M* are pairwise migration rates numbered from population 0 (Africa2) to population 4 (America)

M01	M10	M12	M21	M23	M32	M34	M43	M02	M20	M03	M30	M04	M40	M13	M31	M14	M41	M24	M42
2.1092 e-07	2.3405 e-06	6.8818 e-07	1.0032 e-05	3.7624 e-06	2.0730 e-06	3.9028 e-06	4.6873 e-06	1.0517 e-06	1.6399 e-06	1.4843 e-07	1.1139 e-06	1.3739 e-06	2.9639 e-07	3.4169 e-07	1.4061 e-07	2.6611 e-07	3.3381 e-07	3.7917 e-06	2.7227 e-06
4.0924 e-07	1.3297 e-07	9.7725 e-07	1.0824 e-06	5.4905 e-07	1.6803 e-06	1.5580 e-06	1.3040 e-06	2.6228 e-07	1.9271 e-07	8.7827 e-07	1.6083 e-07	4.3319 e-07	1.2945 e-06	4.1117 e-07	1.0548 e-06	7.6409 e-07	1.7891 e-06	1.3625 e-06	5.1306 e-07
7.7355 e-06	1.2793 e-05	8.4155 e-06	1.1639 e-05	9.4577 e-06	9.0864 e-06	8.1415 e-06	8.7104 e-06	7.0558 e-06	1.1664 e-05	8.5424 e-06	1.3895 e-05	8.1077 e-06	8.6616 e-06	8.5431 e-06	7.7131 e-06	8.0440 e-06	8.657 8e-06	8.6370 e-06	8.8781 e-06

this case, the number of generations per year would be ~4.8 and the other times to bacterial population splits (T2, T3, and T4) would translate into much more recent events, although the relative timing of colonization of different geographic regions in absolute number of generations would not be affected.

H. pylori generation time is thus a key parameter in the estimation of coevolutionary times of host–parasite population differentiation and also to make a comparison between our inferred long-term mutation rate with previous estimates, which are calibrated in years instead of generations. Although two generations per year may seem unreasonably slow for a bacterial organism, we cannot exclude that the peculiar epidemiological dynamics of this bacterium, such as lifelong infection and acquisition early in life (see Introduction), may influence the long term generation time here considered. Both experimental (*i.e.*, familial studies of age structured host samples) and analytical epidemiological models could be used to obtain an empirical estimate. Since *H. pylori* strains could not have colonized any area before the arrival of their human host, our proposed generational time can be considered a lower limit.

Apart from methodological limitations, the events and their timings elucidated here are largely congruent with the human genetic and archaeological literature, confirming previous hypotheses of a close coevolutionary relationship between the two species (Linz *et al.* 2007; Moodley *et al.* 2012). The divergence of the African strains associated with the San, assuming a good fit between human and bacterial estimates, supports an ancient origin of human *Helicobacter*—seeming to have been already in association with the human host before the separation of the San population, and older than an association of at least ~100,000 suggested by MLST sequences (Moodley *et al.* 2012). Given the high level of host specialization, one may hypothesize that this stomach pathogen evolved along with the human host early in the genus *Homo*—a model of interest for future investigation.

Most interestingly, from the bacterial perspective, are the strong signals of population size reduction within Africa, particularly dramatic in the case of the San-associated Africa2. This could have resulted from a reduction in the effective size of the human host population itself, as we know that San hunter-gatherer populations were adversely affected by the Bantu expansion (>1000 yr ago) and by more recent European colonization. However, this does not explain a similar but not as strongly negative growth rate in Africa1 strains, associated with the Bantu and other African populations, which are known to have increased in population effective size since the Neolithic revolution. One alternative to human demography may be stronger selection in Africa, a notion that is consistent with the larger number of putatively adaptive regions identified in Africa, relative to other sampled populations (Figure 3 and Table S3). Despite the very high prevalence of *H. pylori* on this continent, a significant association with

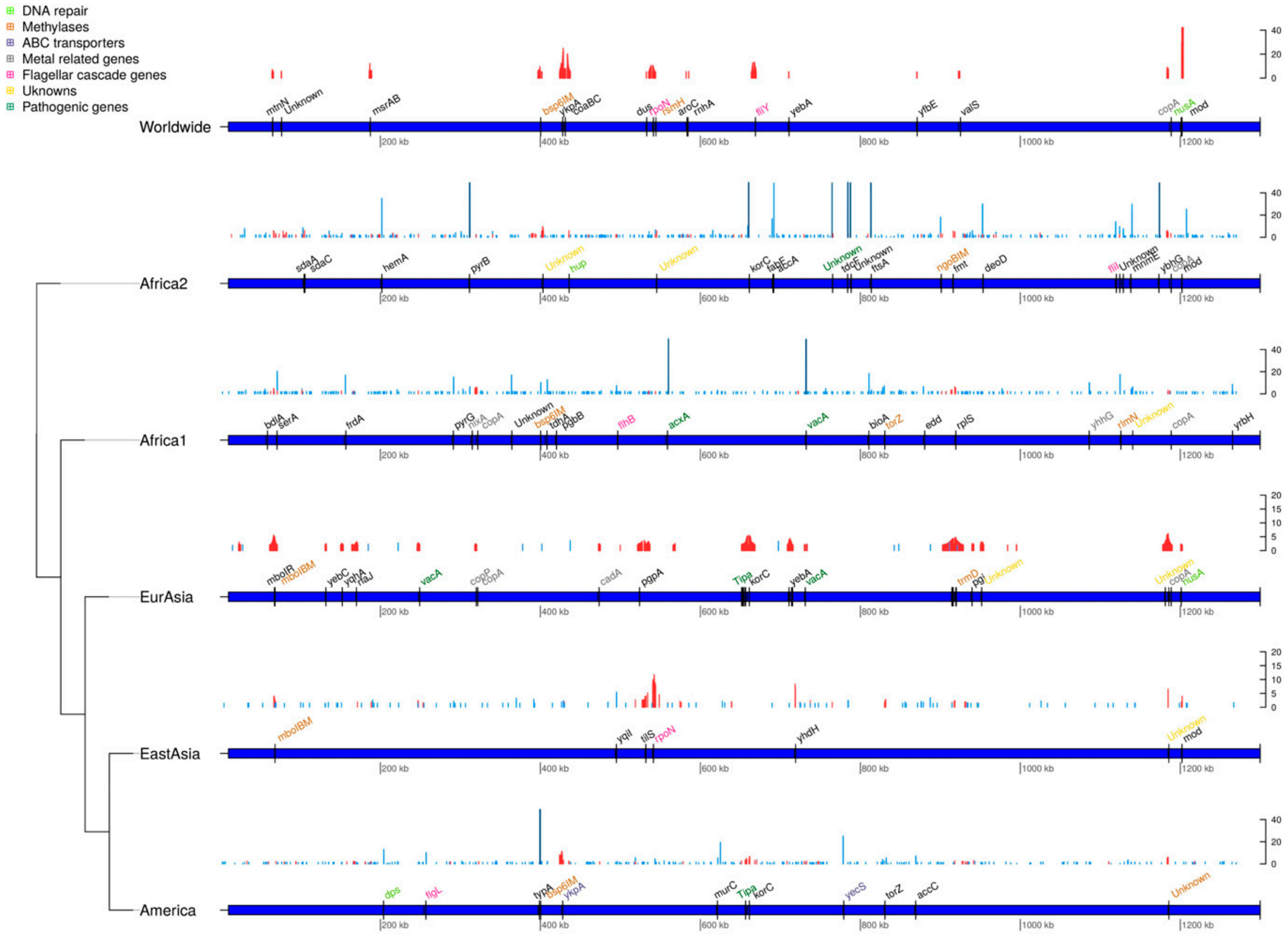


Figure 3 Results of the SweeD and OmegaPlus analyses. A comparative representation for a "synthetic" strain of the worldwide sample and one synthetic strain of each population is drawn using a fictitious topology. Selection values are reported on the graph above each synthetic strain, on the *y*-axis, and genomic position on the *x*-axis. Omega values are represented with red lines, while alpha values are reported in blue. Since alpha values reach much higher levels than omega values, to make the figure easy to read, we reported both omega and alpha values within a scale from zero to 50, and we indicated alpha values >50 in darker blue. Genes falling into the functional categories explained in the discussion are color coded as reported in the legend, while remaining are in black. This figure was generated using R bioconductor package *genoPlotR* (Guy *et al.* 2010).

the incidence of gastric diseases has never been demonstrated (Graham *et al.* 2007; Bauer and Meyer 2011). The opposite is true in non-African strains, where we show that *H. pylori* had a very low ancestral effective population size, coupled with the high population growth rates in our global sample. It may, therefore, be reasonable to hypothesize that the long-term African association of this bacterium with human populations may have led to selection for reduced pathogenicity, whereas a founder effect and rapid growth during the colonization of populations in other areas of the world could have freed this population from these long-term selective constraints, possibly resulting in a more virulent and pathogenic bacterial population (Argent *et al.* 2008; Duncan *et al.* 2013). Concerning the divergence of the American population, we did not detect a clear signature of a founder event. Although the timing of the population split fits with the estimated human colonization of the Americas, we acknowledge an important lack of sampling coverage of the vast Siberia region, which hinders more conclusive results on the expansion dynamics of *H. pylori* across East Asia and to the Americas.

The results obtained with different selection methods address somewhat different biological questions and the extent to which each of these is robust to nonequilibrium demographic histories has only been partially described (*e.g.*, Crisci *et al.* 2013). Based on our inferred demographic history, however, it is possible to describe the true and false positive rates of these statistics for our specific model of interest—representing an empirical solution that may partially overcome such limitations. Global signatures of selection were found in association with several genes of unknown function.

Worldwide and population-specific genes under selection

Patterns of local adaptation are potentially of great medical interest, as they may help explain the continentally differing patterns of virulence observed thus far (Wroblewski *et al.*

Table 4 List of genes identified as being under positive selection by population, classified by function

			Population					
Functional group and genes			Wd	Af2	Af1	EuAs	EaAs	Am
DNA repair								
nusA	1200135–1201322	Transcription elongation protein nusA	*			*		
hup	436148–435789	DNA-binding protein HU		*				
dps	204062–203625	DNA protection during starvation protein						*
Methylases								
vspIM	394565–397009	Modification methylase VspI	*					
bsp6IM	400156–400575	Modification methylase Bsp6I	*		*			*
rimO	525847–527163	Ribosomal protein S12 methylthiotransferase RimO				*		
rsmH	544664–543756	Ribosomal RNA small subunit methyltransferase H	*				*	
mboIBM	67666–68421	Modification methylase *Mbo*IB				*	*	
torZ	830889–830026	Trimethylamine-N-oxide reductase			*			*
ngoBIM	901339–900317	Modification methylase NgoBI		*				
trmD	920231–919542	tRNA guanine-N1–methyltransferase				*		
rlmN	1125936–1124869	Ribosomal RNA large subunit methyltransferase N			*			
ABC transporters								
ykpA	426574–428175	Uncharacterized ABC transporter ATP-binding protein YkpA						*
yecS	779424–778525	Probable amino-acid ABC transporter permease protein HI_0179						*
Metal-related genes								
copA	321905–319935	Copper-transporting ATPase	*	*	*	*		
copP	319934–319734	Copper-associated protein				*		
copA	1188908–1186560	Copper-exporting P-type ATPase A			*	*		
nixA	316216–317211	High-affinity nickel-transport protein nixA			*			
yhhG	1086396–1085875	Putative nickel-responsive regulator			*			
cadA	471626–473686	Cadmium zinc and cobalt-transporting ATPase				*		
Falgellar Cascade genes								
fliY	668666–669037	Flagellar motor switch phosphatase FliY	*					
fliI	1120192–1118888	Flagellum-specific ATP synthase		*				
flhB	496895–495987	Flagellar biosynthetic protein flhB			*			
flgL	257020–254537	Flagellar hook-associated protein 3						*
rpoN	540290–541495	RNA polymerase sigma-54 factor	*				*	
Unknowns								
Unknown	401880–403466	Outer membrane protein		*				
Unknown	545671–544850	Outer membrane protein		*				
Unknown	560264–559263	Outer membrane protein			*			
Unknown	951951-950851	Outer membrane protein				*		
Unknown	1138544–1140799	Outer membrane protein			*			
Unknown	1185912–1184782	Outer membrane protein				*	*	*
Pathogenic genes								
Tipα	656788–656210	Tumor necrosis factor alpha-inducing protein				*		*
vacA	247650–249185	Vacuolating cytotoxin autotransporter				*		
vacA	731291–732442	Vacuolating cytotoxin autotransporter			*	*		
Unknown	764984–765604	Cytotoxin-protein like vacA		*				
acxA	559085–556944	Acetone carboxylase beta subunit			*			

Populations are abbreviated as Wd, worldwide; Af2, Africa2; Af1, Africa1; EuAs, EuroAsia; Eas, EastAsia; Am, America. For a complete list of genes identified as being putatively positively selected in worldwide and local samples see Table S3.

2010; Bauer and Meyer 2011; Matsunari *et al.* 2012; Shiota *et al.* 2013). TheAfrica2 population shows the strongest evidence of recurrent local adaptation, a result that is perhaps intuitive given its long association with the San, one of the most ancient of human groups. Adaptive events within Africa2 include the protein-coding *ftsA* gene (Table S3), which is associated with the cytoskeletal assembly during bacterial cell division (Loose and Mitchison 2014). In addition, results from the analysis of the Africa1 population highlight potentially interesting aspects of the long-term adaptation of *H. pylori* to this population. Among European strains, we identified the only instance in which an antibiotic-associated gene (the penicillin binding protein 1A, *mrcA*) (Table S3) was under selection. This gene was experimentally shown to confer resistance to β-Lactam when a single amino acid substitution occurs (Ser414→Arg) (Gerrits *et al.* 2002). Although our annotated genome of the EuroAsian strains does not show this specific alteration, European *H. pylori* has been more likely exposed to antibiotic treatments than in other regions of the world. On the other hand, recent positive selection at the global level as a consequence of the use of antibiotics seems unlikely, as antibiotic treatment has not been implemented on a global scale. Surprisingly, our analysis did not detect relevant signatures of selection among EastAsian strains, despite the well-known medical risk of gastric cancer associated with these strains. The

American population showed the strongest signature of selection associated with a GTP binding protein whose role is still unknown (*typA*) (Table S3). Our overall results concerning putatively positively selected genes support the role of important metabolic pathways associated with structural and motility functions. This study thus highlights important candidates for future experimental and functional selection studies (for a complete list of candidate genes see Table S4).

Genes involved in DNA repair:

Worldwide genomic regions under selection were identified by OmegaPlus (Table S3), with the strongest signature of selection at the transcription elongation factor gene *nusA*, also flagged locally among EuroAsian strains. In *E. coli*, this protein plays an important role in DNA repair and damage tolerance (Cohen *et al.* 2010). Since *H. pylori* infection of human stomachs can compromise host-cell integrity, inducing breaks in the double-strand and a subsequent DNA damage response (Toller *et al.* 2011), an efficient DNA repair mechanism could be important in protecting bacterial DNA from damage induced by itself or in response to altered physiological conditions in the host stomach. Along with this, indications of positive selection for genes protecting DNA integrity were found among Africa2 and American strains: HU binding protein (*hup*) and during starvation protein (*dps*), respectively (Table 4). The former protein protects DNA from stress damage in *H. pylori* (Wang *et al.* 2012), while the latter is required for survival during acid stress, although its role has been characterized in *E. coli* but not in *H. pylori* (Jeong *et al.* 2008).

Genes involved in methylation patterns:

Several genes expressing proteins involved in DNA methylation were identified as likely under selection (Table 4). A recent study by Furuta *et al.* (2014) used a genomic approach to compare methylation profiles of closely related *H. pylori* strains and showed outstanding diversity of methylation sequence-specificity across lineages. As methylation is an epigenetic mechanism responsible for the regulation of gene expression and phenotypic plasticity, the identification of certain selected methylation genes encourage the study of their specific role and their evolutionary implications in *H. pylori* methylation patterns.

ATP binding cassette transporters:

The ATP binding cassette (ABC) transporters are ubiquitous, and among their functions is the ability to expel cytotoxic molecules out of the cell, conferring resistance to drugs (Linton 2007). Two of these uncharacterized genes were indicated to be under positive selection in American strains (*ykpA* and *yecS*) Table 4).

Genes involved in flagellar cascade:

Cell motility and cell adherence to the stomach mucosa is a key factor for the successful colonization of the human stomach, and several positively selected flagellum-specific genes (*flgL*, *flhB*, *fliI*, and *fliY*) were identified across different local populations (Table 4). Apart from genes involved into the flagellar cascade, positive selection was also detected in the regulating factor of the cascade itself (sigma(54) or *rpoN*), corroborating the importance of bacterial motility in survival (Table 4).

Genes involved in heavy metal metabolism:

Importantly, our selection analysis highlights a potentially predominant role for genes associated with copper metabolism in the *H. pylori* life cycle, with the same genes flagged in multiple populations (Table 4). Copper-mediated colonization of the stomach mucosa occurs through the action of trefoil peptides in *H. pylori* (Montefusco *et al.* 2013) and copper drastically increases in cancerous tissues. However the detailed role of *copA* and *copP* genes and of copper metabolism in *H. pylori* long-term adaptation is yet to be investigated. Interestingly, the Africa1 population shows signatures of positive selection of two genes involved in the transport and regulation of nickel (*nixA* and *yhhG*), while EuroAsian strains show hints of selection for a cadium, zinc, and cobalt transporter (*cadA*) (see Table 4).

Genes involved in virulence:

We identified a number of putatively selected *vacA* genes in local populations as expected from previous indications of their importance in *H. pylori* pathogenicity (Olbermann *et al.* 2010). It is further interesting to note that *vacA* and *vacA-like* genes also show evidence for selection among African populations, where the association of *H. pylori* with gastric disease is not considered to be significant. In particular, Africa1 strains present a strong signal associated with the acetone carboxylase beta subunit (*acxA*) (Table 4), which is part of the pathologically relevant operon *acxABC*, as it is associated with virulence and survival of the bacterium into the host stomach (Brahmachary *et al.* 2008; Risch 2012). These observations suggest that virulence-related genes may nonetheless play an important role in bacterial adaptation or, more specifically, that *H. pylori* may indeed have a pathogenic role among African populations that is masked by other factors leading to gastric diseases. Finally, the tumor necrosis factor alpha-inducing protein (*Tipα*) (Suganuma *et al.* 2001, 2006, 2008) was identified in EuroAsian and American strains, calling for closer investigation in relation to its potentially pathogenic role among these specific populations.

Outer membrane proteins:

Many unknown genes appear in the list of putatively selected genes (Table 4). Among those, there could be a particular interest in further investigating the nature and role of outer membrane proteins (OMPs), which would certainly provide valuable information on the interaction of *H. pylori* and the gastric environment. There are at least five recognized families of genes coding for OMP (HopA–E), which are involved in the processes of adherence to the gastric mucosa and thus play an important role in successful colonization of the host's stomach (Yamaoka 2008; Oleastro and Ménard 2013). Moreover, the importance of specific OMP genes in H. pylori has been investigated in recent studies (Kennemann *et al.* 2012; Nell *et al.* 2014).

From an evolutionary perspective, our study presents evidence for processes of adaptation in H. pylori to its human host, but, regrettably, does not provide a perspective on the coevolutionary interactions that are likely to have occurred during their long history of association. In this sense, it is intriguing to speculate that the interaction with the human host did not simply lead to pathogenic conditions but also led to mutual adaptation. Theories on beneficial interactions of *H. pylori* and the human host have been already suggested (Blaser 2008). The observation that <15% of infected human individuals show clinical symptoms has led previous studies to speculate that *H. pylori* may play an important, but not necessarily pathogenic, role in the human gastric niche, potentially even protecting its host from other gastric infections (Blaser 2008; Shahabi *et al.* 2008; Atherton and Blaser 2009). In support of this idea, a recent survey among native Americans reported that patients with lower host–bacteria coancestry—that is, patients infected with hpEurope (here included into the Eurasian population) and not with hspAmerind (the American population)—show increased severity of premalignant lesions in gastric cancer (Kodaman *et al.* 2014). Hopefully, future investigation will also focus on the long-term interaction of the two species and the possible signatures in the human genome that result from the long association with *H. pylori*.

Although our results highlighting major selective events in Africa are supported by a common African origin for both species, the coevolutionary history between *H. pylori* and humans is an area that warrants future and more detailed investigation at the genomic level. A first step would be the inclusion of more genomes from underrepresented regions such as Sahul, North-East Africa, Central Asia, and the Americas. Furthermore, unrepresented regions such as Siberia and Oceania would allow for the investigation of genetic continuity/discontinuity across northeastern and southeastern Asia to the Americas and the Pacific, respectively. A deeper analysis of Asian, American, and Austronesian bacterial genomes may also help shed light on alternative Pacific routes for the colonization of the Americas, a hypothesis that has been widely debated in the literature (see Gonçalves *et al.* 2013; Malaspinas *et al.* 2014).

Acknowledgments

We thank Mark Achtman for useful comments on an early version of the manuscript and the Vital-IT bioinformatic center for the technical support with the Vital-IT cluster. This project was supported by European Research Area Net PathoGenoMics project Analysis of genetic diveristy of Helicobacter HELDIVNET (0313930B) from the German Ministry of Education and Research (BMBF) to S.S. and BMBF Project 01GS0805 to R.R. for massive parallel sequencing. V.M. was supported by a postdoctoral fellowship from the European Union (Framework 7) and a short-term postdoctoral fellowship from the European Molecular Biology Organization. J.D.J. and M.F. were supported by grants from the Swiss National Science Foundation and a European Research Council starting grant (to J.D.J.). X.D. acknowledges the National Institute for Health Research for Health Protection Research Unit funding. The authors declare no conflict of interest.

Literature Cited

Alachiotis, N., A. Stamatakis, and P. Pavlidis, 2012 OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics 28: 2274–2275.

Allaker, R. P., K. A. Young, J. M. Hardie, P. Domizio, and N. J. Meadows, 2002 Prevalence of helicobacter pylori at oral and gastrointestinal sites in children: evidence for possible oral-to-oral transmission. J. Med. Microbiol. 51: 312–317.

Argent, R. H., J. L. Hale, E. M. El-Omar, and J. C. Atherton, 2008 Differences in Helicobacter pylori CagA tyrosine phosphorylation motif patterns between western and East Asian strains, and influences on interleukin-8 secretion. J. Med. Microbiol. 57: 1062–1067.

Atherton, J. C., and M. J. Blaser, 2009 Coadaptation of Helicobacter pylori and humans: ancient history, modern implications. J. Clin. Invest. 119: 2475–2487.

Basso, D., C.-F. Zambon, D. P. Letley, A. Stranges, A. Marchet *et al.*, 2008 Clinical relevance of Helicobacter pylori cagA and vacA gene polymorphisms. Gastroenterology 135: 91–99.

Bauer, B., and T. F. Meyer, 2011 The human gastric pathogen *Helicobacter pylori* and its association with gastric cancer and ulcer disease. Ulcers 10.1155/2011/340157.

Beier, D., G. Spohn, R. Rappuoli, and V. Scarlato, 1997 Identification and characterization of an operon of Helicobacter pylori that is involved in motility and stress adaptation. J. Bacteriol. 179: 4676–4683.

Blaser, M. J., 2008 Disappearing microbiota: Helicobacter pylori protection against esophageal adenocarcinoma. Cancer Prev. Res. (Phila.) 1: 308–311.

Brahmachary, P., G. Wang, S. L. Benoit S, M. V. Weinberg, R. J. Maier *et al.*, 2008 The human gastric pathogen Helicobacter pylori has a potential acetone carboxylase that enhances its ability to colonize mice. BMC Microbiology 8: 14.

Breurec, S., B. Guillard, S. Hem, S. Brisse, F. B. Dieye *et al.*, 2011 Evolutionary history of Helicobacter pylori sequences reflect past human migrations in Southeast Asia. PLoS ONE 6: e22058.

Brown, L. M., 2000 Helicobacter pylori: epidemiology and routes of transmission. Epidemiol. Rev. 22: 283–297.

Bures, J., M. Kopácová, I. Koupil, V. Vorísek, S. Rejchrt *et al.*, 2006 Epidemiology of Helicobacter pylori infection in the Czech Republic. Helicobacter 11: 56–65.

Burns, B. P., S. L. Hazell, G. L. Mendz, T. Kolesnikow, D. Tillet *et al.*, 2000 The Helicobacter pylori pyrB gene encoding aspartate carbamoyltransferase is essential for bacterial survival. Arch. Biochem. Biophys. 380: 78–84.

Burnie, J. P., R. C. Matthews, T. Carter, E. Beaulieu, M. Donohoe *et al.*, 2000 Identification of an immunodominant ABC transporter in methicillin-resistant Staphylococcus aureus infections. Infect. Immun. 68: 3200–3209.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza, 1994 *The History and Geography of Human Genes*, Princeton University Press, Princeton, NJ.

Chen, J., X. L. Bu, Q. Y. Wang, P. J. Hu, and M. H. Chen, 2007 Decreasing seroprevalence of Helicobacter pylori infection during 1993–2003 in Guangzhou, Southern China. Helicobacter 12: 164–169.

Cohen, S. E., C. A. Lewis, R. A. Mooney, M. A. Kohanski, J. J. Collins *et al.*, 2010 Roles for the transcription elongation factor NusA in both DNA repair and damage tolerance pathways in Escherichia coli. Proc. Natl. Acad. Sci. USA 107: 15517–15522.

Corander, J., and P. Marttinen, 2006 Bayesian identification of admixture events using multilocus molecular markers. Mol. Ecol. 15: 2833–2843.

Corander, J., P. Marttinen, J. Sirén, and J. Tang, 2008 Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9: 539.

Crisci, J. L., Y.-P. Poh, S. Mahajan, and J. D. Jensen, 2013 The impact of equilibrium assumptions on tests of selection. Front. Genet. 4: 235.

Darling, A. C. E., B. Mau, F. R. Blattner, and N. T. Perna, 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14: 1394–1403.

Darling, A. E., B. Mau, and N. T. Perna, 2010 progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5: e11147.

Didelot, X., and D. Falush, 2007 Inference of bacterial microevolution using multilocus sequence data. Genetics 175: 1251–1266.

Didelot, X., S. Nell, I. Yang, S. Woltemate, S. van der Merwe *et al.*, 2013 Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc. Natl. Acad. Sci. USA 110: 13880–13885.

Van Domselaar, G. H., P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo *et al.*, 2005 BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 33: W455–W459.

Duncan, S. S., P. L. Valk, M. S. McClain, C. L. Shaffer, J. A. Metcalf *et al.*, 2013 Comparative genomic analysis of East Asian and Non-Asian Helicobacter pylori strains identifies rapidly evolving genes. PLoS ONE 8: e55120.

van Duynhoven, Y. T., and R. de Jonge, 2001 Transmission of Helicobacter pylori: A role for food? Bull. World Health Organ. 79: 455–460.

Eriksson, A., L. Betti, A. D. Friend, S. J. Lycett, J. S. Singarayer *et al.*, 2012 Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc. Natl. Acad. Sci. USA 109: 16089–16094.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. 9: e1003905.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox *et al.*, 2001 Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. USA 98: 15056–15061.

Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens *et al.*, 2003 Traces of human migrations in Helicobacter pylori populations. Science 299: 1582–1585.

Festa, R. A., and D. J. Thiele, 2012 Copper at the front line of the host-pathogen battle. PLoS Pathog. 8: e1002887.

Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977–993.

Furuta, Y., H. Namba-Fukuyo, T. F. Shibata, T. Nishiyama, S. Shigenobu *et al.*, 2014 Methylome diversification through changes in DNA methyltransferase sequence specificity. PLoS Genet. 10: e1004272.

Gerrits, M. M., D. Schuijffel, A. A. van Zwet, E. J. Kuipers, C. M. J. E. Vandenbroucke-Grauls *et al.*, 2002 Alterations in penicillin-binding protein 1A confer resistance to β-lactam antibiotics in Helicobacter pylori. Antimicrob. Agents Chemother. 46: 2229–2233.

Ghose, C., G. I. Perez-Perez, L. J. van Doorn, M. G. Domínguez-Bello, and M. J. Blaser, 2005 High frequency of gastric colonization with multiple Helicobacter pylori strains in Venezuelan subjects. J. Clin. Microbiol. 43: 2635–2641.

Gonçalves, V. F., J. Stenderup, C. Rodrigues-Carvalho, H. P. Silva, H. Gonçalves-Dornelas *et al.*, 2013 Identification of Polynesian mtDNA haplogroups in remains of Botocudo Amerindians from Brazil. Proc. Natl. Acad. Sci. USA 110: 6465–6469.

Graham, D. Y., Y. Yamaoka, and H. M. Malaty, 2007 Thoughts about populations with unexpected low prevalences of *Helicobacter pylori* infection. Trans. R. Soc. Trop. Med. Hyg. 101: 849–851.

Gressmann, H., B. Linz, R. Ghai, K. P. Pleissner, R. Schlapbach *et al.*, 2005 Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. PLoS Genet. 1: e43.

Guy, L., J. Roat Kultima, and S. G. E. Andersson, 2010 genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26: 2334–2335.

Hammer, M. F., A. E. Woerner, F. L. Mendez, J. C. Watkins, and J. D. Wall, 2011 Genetic evidence for archaic admixture in Africa. Proc. Natl. Acad. Sci. USA 108: 15123–15128.

Jensen, J. D., K. R. Thornton, C. D. Bustamante, and C. F. Aquadro, 2007a On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176: 2371–2379.

Jensen, J. D., V. L. Bauer DuMont, A. B. Ashmore, A. Gutierrez, and C. F. Aquadro, 2007b Patterns of sequence variability and divergence at the diminutive gene region of Drosophila melanogaster: complex patterns suggest an ancestral selective sweep. Genetics 177: 1071–1085.

Jeong, K. C., K. F. Hung, D. J. Baumler, J. J. Byrd, and C. W. Kaspar, 2008 Acid stress damage of DNA is prevented by Dps binding in Escherichia coli O157:H7. BMC Microbiol. 8: 181.

Jombart, T., 2008 adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.

Jombart, T., and I. Ahmed, 2011 adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11: 94.

Kawai, M., Y. Furuta, K. Yahara, T. Tsuru, K. Oshima *et al.*, 2011 Evolution in an oncogenic bacterial species with extreme genome plasticity: Helicobacter pylori East Asian genomes. BMC Microbiol. 11: 104.

Kennemann, L., X. Didelot, T. Aebischer, S. Kuhn, B. Drescher *et al.*, 2011 Helicobacter pylori genome evolution during human infection. Proc. Natl. Acad. Sci. USA 108: 5033–5038.

Kennemann, L., B. Brenneke, S. Andres, L. Engstrand, and T. F. Meyer *et al.*, 2012 In vivo sequence variation in HopZ, a phase-variable outer membrane protein of Helicobacter pylori. Infect. Immun. 80: 4364–4373.

Khalifa, M. M., R. R. Sharaf, and R. K. Aziz, 2010 Helicobacter pylori: A poor man's gut pathogen? Gut Pathog 2: 2.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524.

Kodaman, N., A. Pazos, B. G. Schneider, M. B. Piazuelo, R. Mera *et al.*, 2014 Human and Helicobacter pylori coevolution shapes the risk of gastric disease. Proc. Natl. Acad. Sci. USA 111: 1455–1460.

Krebes, J., X. Didelot, L. Kennemann, and S. Suerbaum, 2014 Bidirectional genomic exchange between Helicobacter pylori strains from a family in Coventry, United Kingdom. Int. J. Med. Microbiol. 304: 1135–1146.

Lawson, D. J., 2013 Populations in statistical genetic modelling and inference. arXiv:1306.0701 [q-bio].

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genet. 8: e1002453.

Linton, K. J., 2007 Structure and function of ABC transporters. Physiology (Bethesda) 22: 122–130.

Linz, B., F. Balloux, Y. Moodley, A. Manica, H. Liu *et al.*, 2007 An African origin for the intimate association between humans and Helicobacter pylori. Nature 445: 915–918.

Linz, B., H. M. Windsor, J. P. Gajewski, C. M. Hake, D. I. Drautz *et al.*., 2013 Helicobacter pylori genomic microevolution during naturally occurring transmission between adults. PLoS ONE 8: e82187.

Linz, B., H. M. Windsor, J. J. McGraw, L. M. Hansen, J. P. Gajewski *et al.*., 2014 A mutation burst during the acute phase of Helicobacter pylori infection in humans and rhesus macaques. Nat. Commun. 5.

Loose, M., and T. J. Mitchison, 2014 The bacterial cell division proteins FtsA and ftsA self-organize into dynamic cytoskeletal patterns. Nat. Cell Biol. 16: 38–46.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

Malaspinas, A.-S., O. Lao, H. Schroeder, M. Rasmussen, M. Raghavan *et al.*, 2014 Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. Curr. Biol. 24: R1035–R1037.

Marshall, B. J., and J. R. Warren, 1984 Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 1: 1311–1315.

Matsunari, O., S. Shiota, R. Suzuki, M. Watada, N. Kinjo *et al.*, 2012 Association between Helicobacter pylori virulence factors and gastroduodenal diseases in Okinawa, Japan. J. Clin. Microbiol. 50: 876–883.

Miller, J. R., A. L. Delcher, S. Koren, E. Venter, B. P. Walenz *et al.*, 2008 Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24: 2818–2824.

Moodley, Y., and B. Linz, 2009 Helicobacter pylori sequences reflect past human migrations. Genome Dyn. 6: 62–74.

Moodley, Y., B. Linz, Y. Yamaoka, H. M. Windsor, S. Breurec *et al.*, 2009 The peopling of the Pacific from a bacterial perspective. Science 323: 527–530.

Moodley, Y., B. Linz, R. P. Bond, M. Nieuwoudt, H. Soodyall *et al.*, 2012 Age of the association between Helicobacter pylori and man. PLoS Pathog. 8: e1002693.

Montefusco, S., R. Esposito, L. D'Andrea, M. C. Monti, C. Dunne *et al.*, 2013 Copper promotes TFF1-mediated Helicobacter pylori colonization. PLoS ONE 8: e79455.

Morelli, G., X. Didelot, B. Kusecek, S. Schwarz, C. Bahlawane *et al.*, 2010 Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. PLoS Genet. 6: e1001036.

Narum, S. R., and J. E. Hess, 2011 Comparison of F(ST) outlier tests for SNP loci under selection. Mol Ecol Resour 11(Suppl 1): 184–194.

Nell, S., D. Eibach, V. Montano, A. Maady, A. Nkwescheu *et al.*, 2013 Recent acquisition of Helicobacter pylori by Baka pygmies. PLoS Genet. 9: e1003775.

Nell, S., L. Kennemann, S. Schwarz, C. Josenhans, and S. Suerbaum, 2014 Dynamics of Lewis b binding and sequence variation of the babA adhesin gene during chronic Helicobacter pylori infection in humans. MBio 5: pii: e02281–e14.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.

Olbermann, P., C. Josenhan, Y. Moodley, M. Uhr, C. Stamer *et al.*, 2010 A global overview of the genetic and functional diversity in the helicobacter pylori cag pathogenicity island. PLoS Genet. 6: e1001069.

Oleastro, M., and A. Menard, 2013 The role of Helicobacter pylori outer membrane proteins in adherence and pathogenesis. Biology (Basel) 2: 1110–1134.

Paradis, E., 2010 pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419–420.

Paradis, E., J. Claude, and K. Strimmer, 2004 APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290.

Patra, R., S. Chattopadhyay, R. De, P. Ghosh, M. Ganguly *et al.*, 2012 Multiple infection and microdiversity among Helicobacter pylori isolates in a single host in India. PLoS ONE 7: e43370.

Pavlidis, P., D. Živkovic, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30: 2224–2234.

Perry, S., M. de la Luz Sanchez, S. Yang, T. D. Haggerty, P. Hurst *et al.*, 2006 Gastroenteritis and transmission of Helicobacter pylori infection in households. Emerg. Infect. Dis. 12: 1701–1708.

Risch, H. A., 2012 Pancreatic cancer: Helicobacter pylori colonization, N-Nitrosamine exposures, and ABO blood group. Mol. Carcinog. 51: 109–118.

Salama, N. R., M. L. Hartung, and A. Müller, 2013 Life in the human stomach: persistence strategies of the bacterial pathogen Helicobacter pylori. Nat. Rev. Microbiol. 11: 385–399.

Salih, B. A., 2009 Helicobacter pylori infection in developing countries: The burden for how long? Saudi J. Gastroenterol. 15: 201–207.

Shiota, S., R. Suzuki, and Y. Yamaoka, 2013 The significance of virulence factors in Helicobacter pylori. J. Dig. Dis. 14: 341–349.

Schlebusch, C. M., P. Skoglund, P. Sjödin, L. M. Gattepaille, D. Hernandez *et al.*, 2012 Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338: 374–379.

Schwarz, S., G. Morelli, B. Kusecek, A. Manica, F. Balloux *et al.*, 2008 Horizontal *vs.* familial transmission of Helicobacter pylori. PLoS Pathog. 4: e1000180.

Shahabi, S., Y. Rasmi, N. H. Jazani, and Z. M. Hassan, 2008 Protective effects of Helicobacter pylori against gastroesophageal reflux disease may be due to a neuroimmunological anti-inflammatory mechanism. Immunol. Cell Biol. 86: 175–178.

Suerbaum, S., and P. Michetti, 2002 Helicobacter pylori infection. N. Engl. J. Med. 347: 1175–1186.

Suerbaum, S., and C. Josenhans, 2007 Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat. Rev. Microbiol. 5: 441–452.

Suganuma, M., M. Kurusu, S. Okabe, N. Sueoka, M. Yoshida *et al.*, 2001 Helicobacter pylori membrane protein 1: a new carcinogenic factor of Helicobacter pylori. Cancer Res. 61: 6356–6359.

Suganuma, M., T. Kuzuhara, K. Yamaguchi, and H. Fujiki, 2006 Carcinogenic role of tumor necrosis factor-alpha inducing protein of Helicobacter pylori in human stomach. J. Biochem. Mol. Biol. 39: 1–8.

Suganuma, M., K. Yamaguchi, Y. Ono, H. Matsumoto, T. Hayashi *et al.*, 2008 TNF-alpha-inducing protein, a carcinogenic factor secreted from H. pylori, enters gastric cancer cells. Int. J. Cancer 123: 117–122.

Sycuro, L. K., T. J. Wyckoff, J. Biboy, P. Born, Z. Pincus *et al.*, 2012 Multiple peptidoglycan modification networks modulate Helicobacter pylori's cell shape, motility, and colonization potential. PLoS Pathog. 8: e1002603.

Tang, J., W. P. Hanage, C. Fraser, and J. Corander, 2009 Identifying currents in the gene pool for bacterial populations using an integrative approach. PLOS Comput. Biol. 5: e1000455.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750.

Toller, I. M., K. J. Neelsen, M. Steger, M. L. Hartung, M. O. Hottiger *et al.*, 2011 Carcinogenic bacterial pathogen Helicobacter pylori triggers DNA double-strand breaks and a DNA damage response in its host cells. Proc. Natl. Acad. Sci. USA 108: 14944–14949.

Veeramah, K. R., D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins *et al.*, 2011 An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol. Biol. Evol. 29: 617–630.

Wang, G., L. F. Lo, and R. J. Maier, 2012 A histone-like protein of Helicobacter pylori protects DNA from stress damage and aids host colonization. DNA Repair (Amst.) 11: 733–740.

Willems, R. J. L., J. Top, W. van Schaik, H. Leavis, M. Bonten *et al.*, Corander J., 2012 Restricted gene flow among hospital subpopulations of Enterococcus faecium. MBio 3: e00151–00112.

Yamaoka, Y. (Editor), 2008 *Helicobacter pylori: Molecular Genetics and Cellular Biology*. Caister Academic Press, Wymondham.

Communicating editor: R. Nielsen

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176404/-/DC1

Worldwide Population Structure, Long-Term Demography, and Local Adaptation of *Helicobacter pylori*

Valeria Montano, Xavier Didelot, Matthieu Foll, Bodo Linz, Richard Reinhardt, Sebastian Suerbaum, Yoshan Moodley, and Jeffrey D. Jensen

DOI: 10.1534/genetics.115.176404

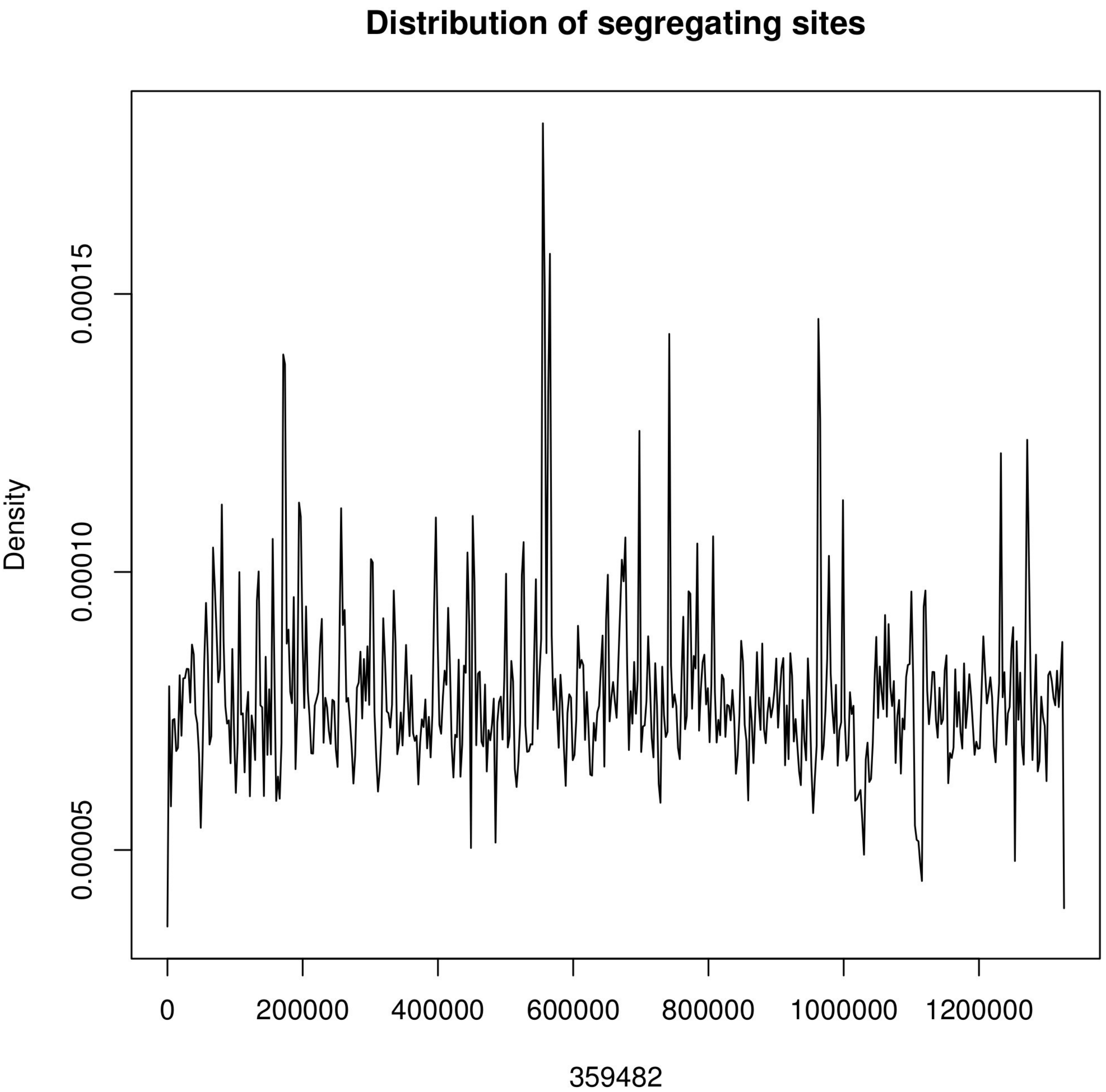
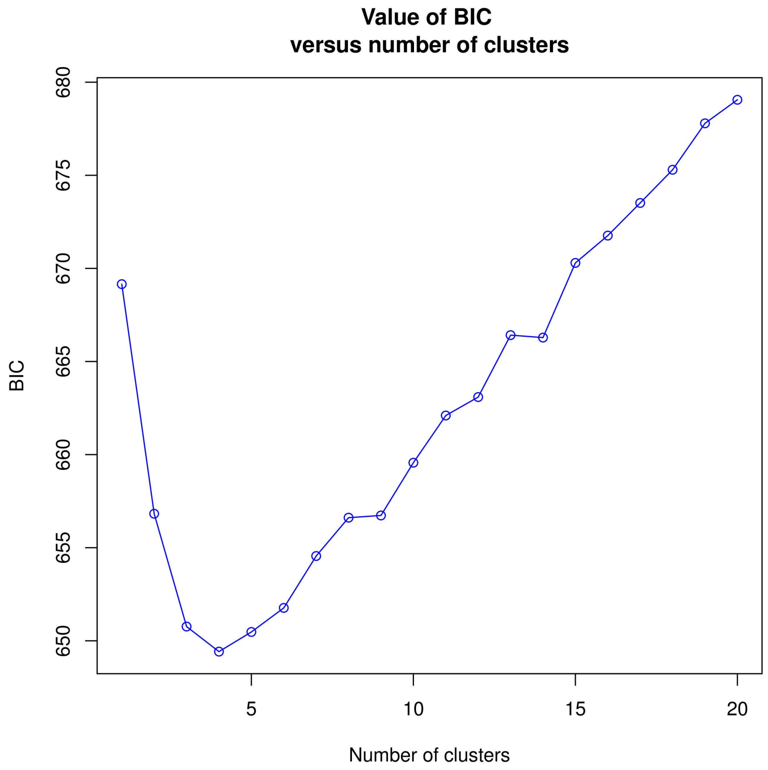


Figure S1. Figure S1. Density distribution of segregating sites along the aligned sequence. The highest peak is observed in the region between 500,000 and 600,000 base pairs.

A)



B)

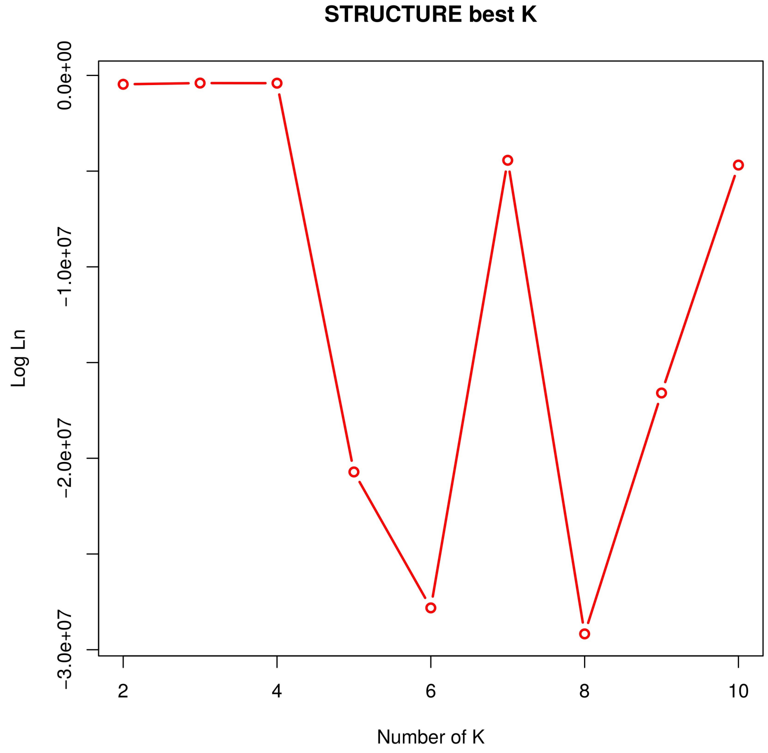


Figure S2. A) Bayesian Information Criterion (BIC) plot from the *find.clusters* function in adegenet package to retrieve the best number of clusters in our sample. The minimum is at *K* = 4 indicating the best point of optimization. B) Plot of structure loglikelihood for each tested *K*, averaged across 5 independent runs per *K*.

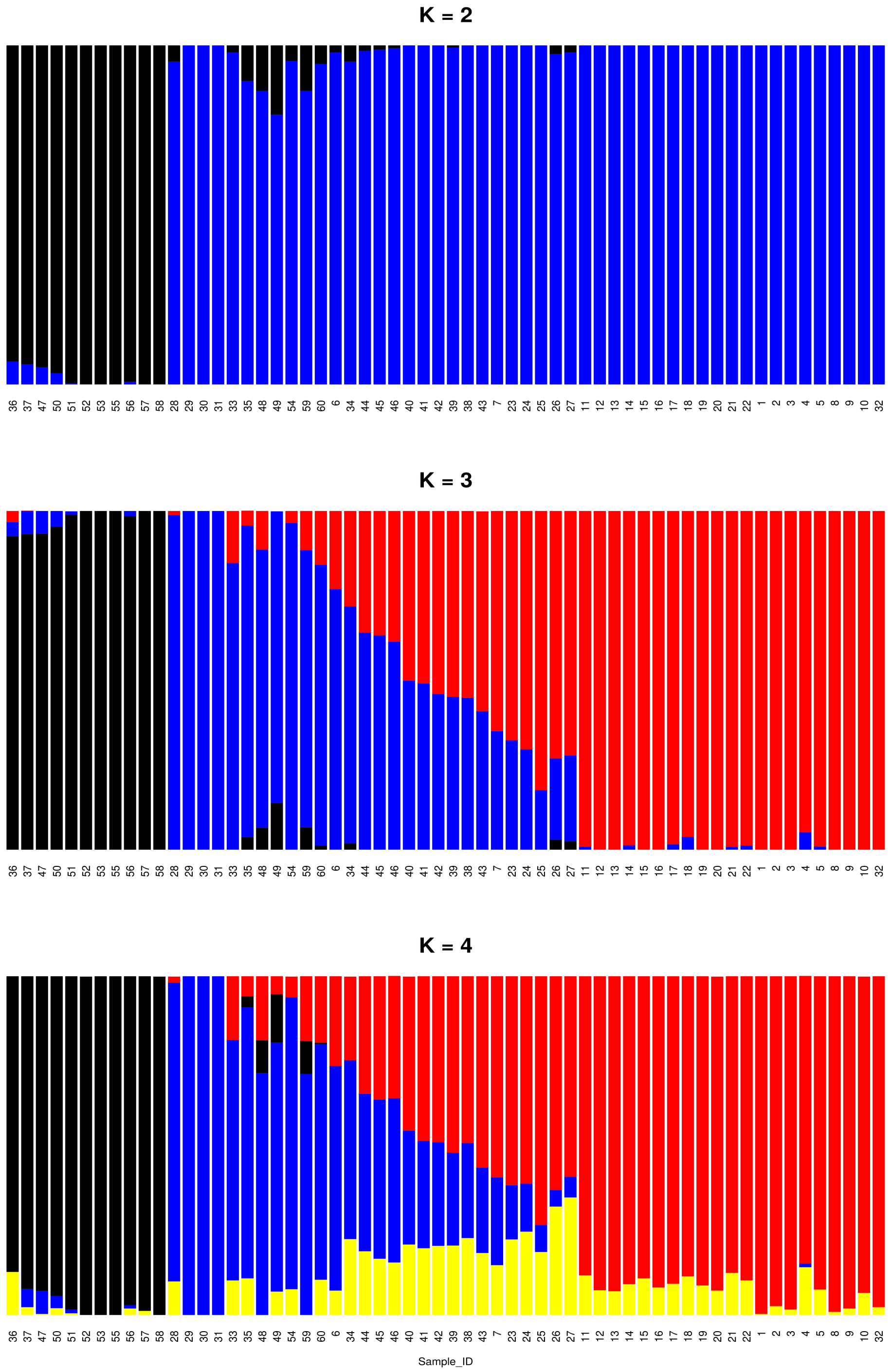


Figure S3. Plot of STRUCTURE analysis with probabilities of individual assignment for *K* from 2 to 4, obtained using the admixture model.

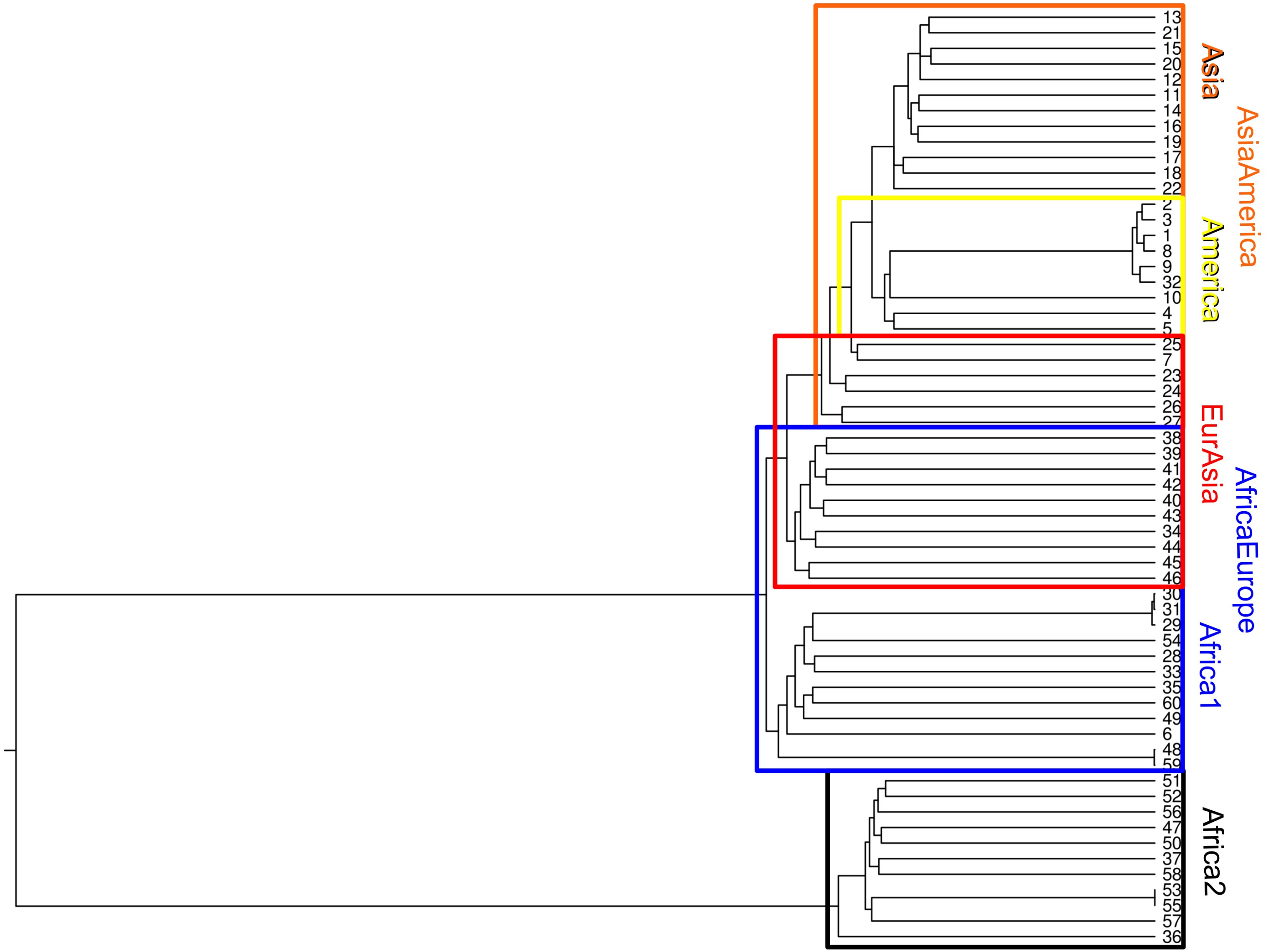


Figure S4 Clonal genealogy of the 60 genomes estimated using ClonalFrame. The DAPC population assignments are indicated with the same colors as in Figure 1.

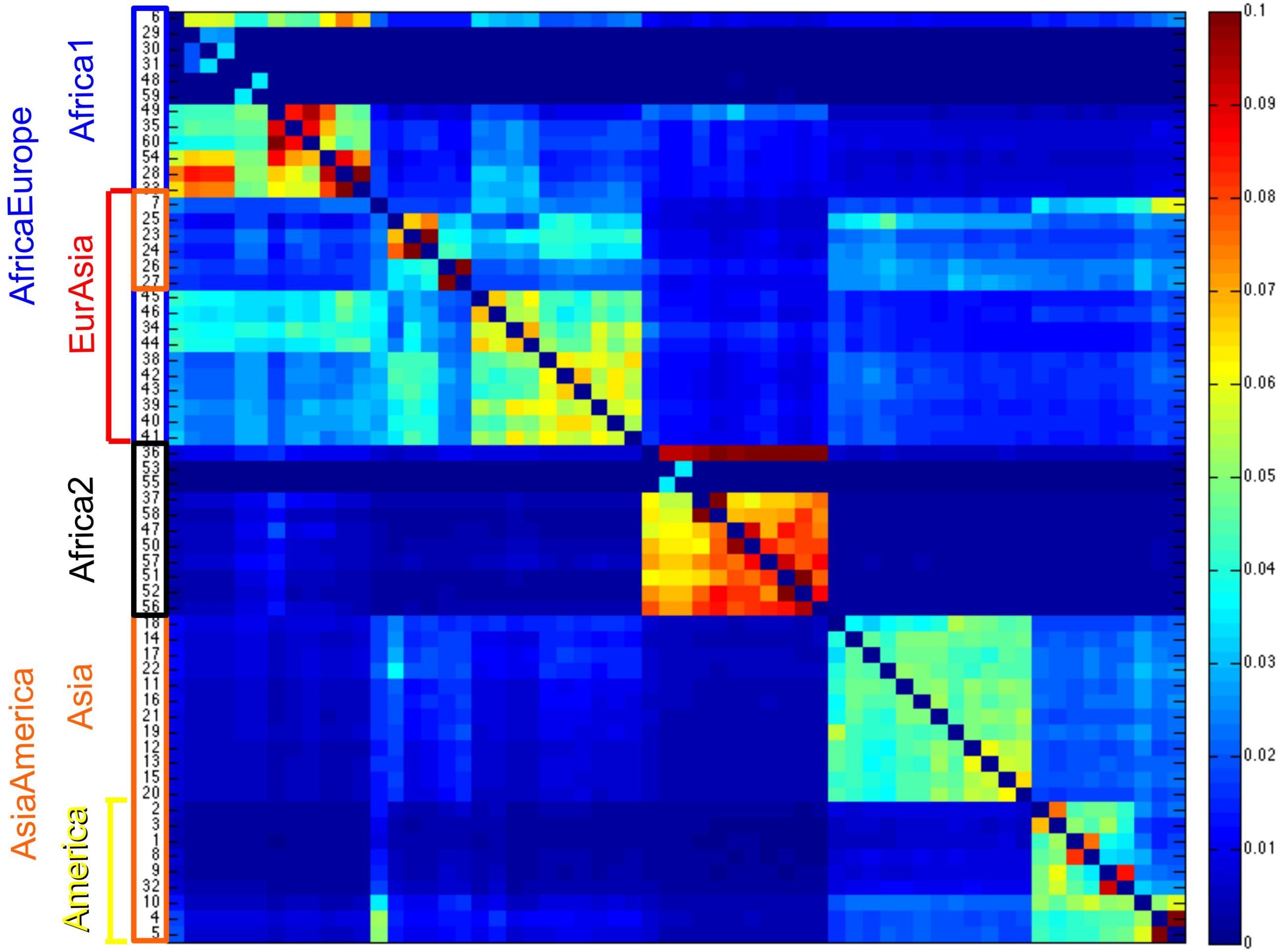


Figure S5 Matrix of co-ancestry among the sequences estimated with fine-STRUCTURE. The hottest tones are associated to higher genetic affinities and coldest tones to lower genetic affinities. The DAPC populations are indicated with colors as in Figure 1.

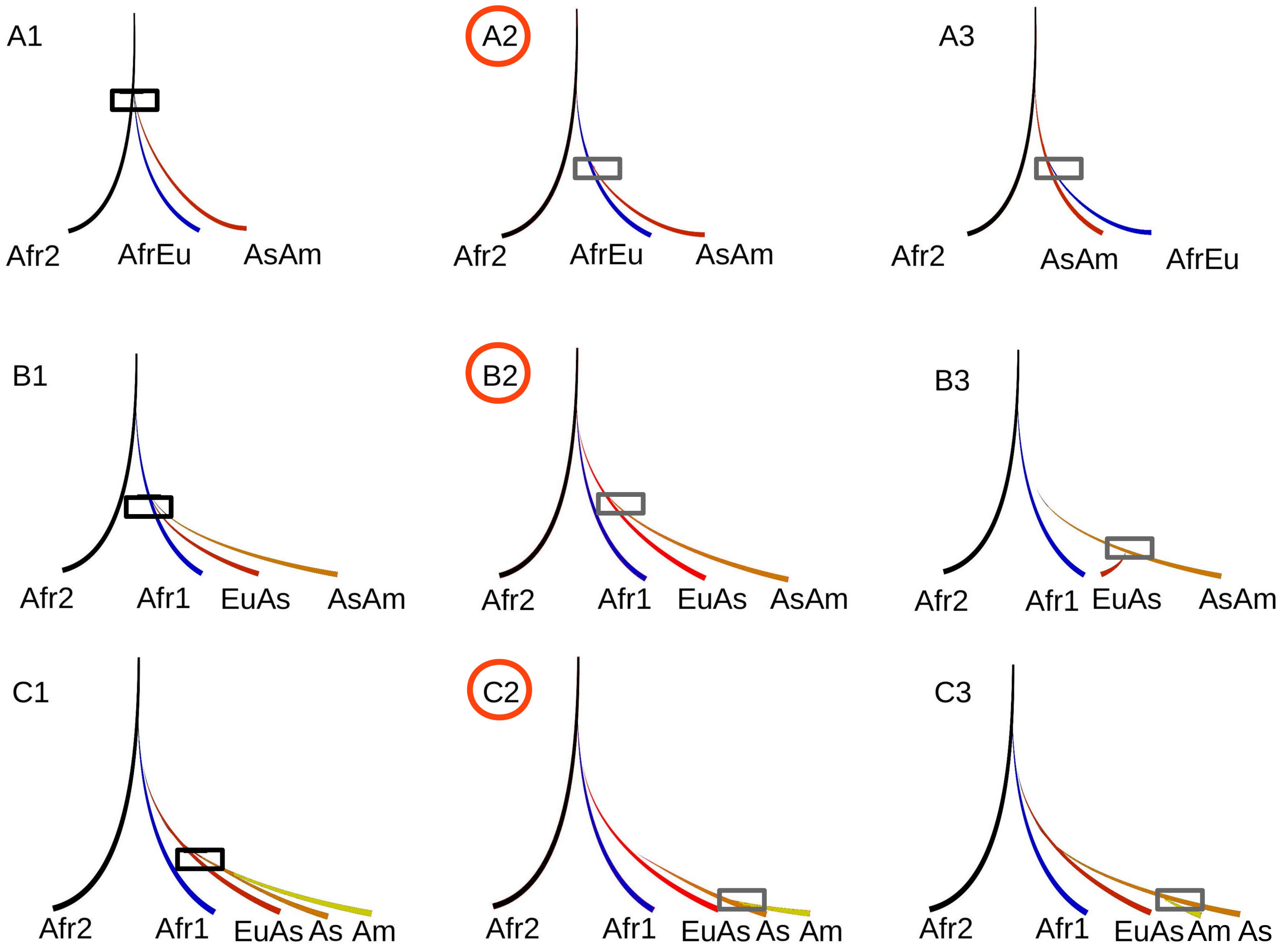


Figure S6 Multiple tree topologies compared under a model of divergence without migration. Populations are color-coded as in Figure 1. Black boxes indicate points of split where two populations diverge independently from the same ancestral pool, while grey boxes indicate points of split where one population diverged from another single population. The best model is marked with a red circle. A) models based on 3 populations: A1) both AfricaEu and AsiaAmerica clusters are derived simultaneously from the ancestral populationAfrica2; A2) Africa2 and AfricaEu first split, followed by a derived AsiaAmerican population; A3) AsiaAmerica population split from Africa2 with AfricaEu derived from it. B) models with 4 populations based on the best genealogy found among previous models: B1) EuAsia and AsiaAmerica are equally derived from Africa1 population at different times; B2) Africa1 and EuAsia populations split in first place and AsiaAmerica population arose afterwards from a split with EuAsia; B3) EuAsia population derived from AsiaAmerica population after a split of the latter from Africa1. C) models of 5 populations, following the same scheme as above in C1, C2 and C3 for sub-clusters Asia and America.

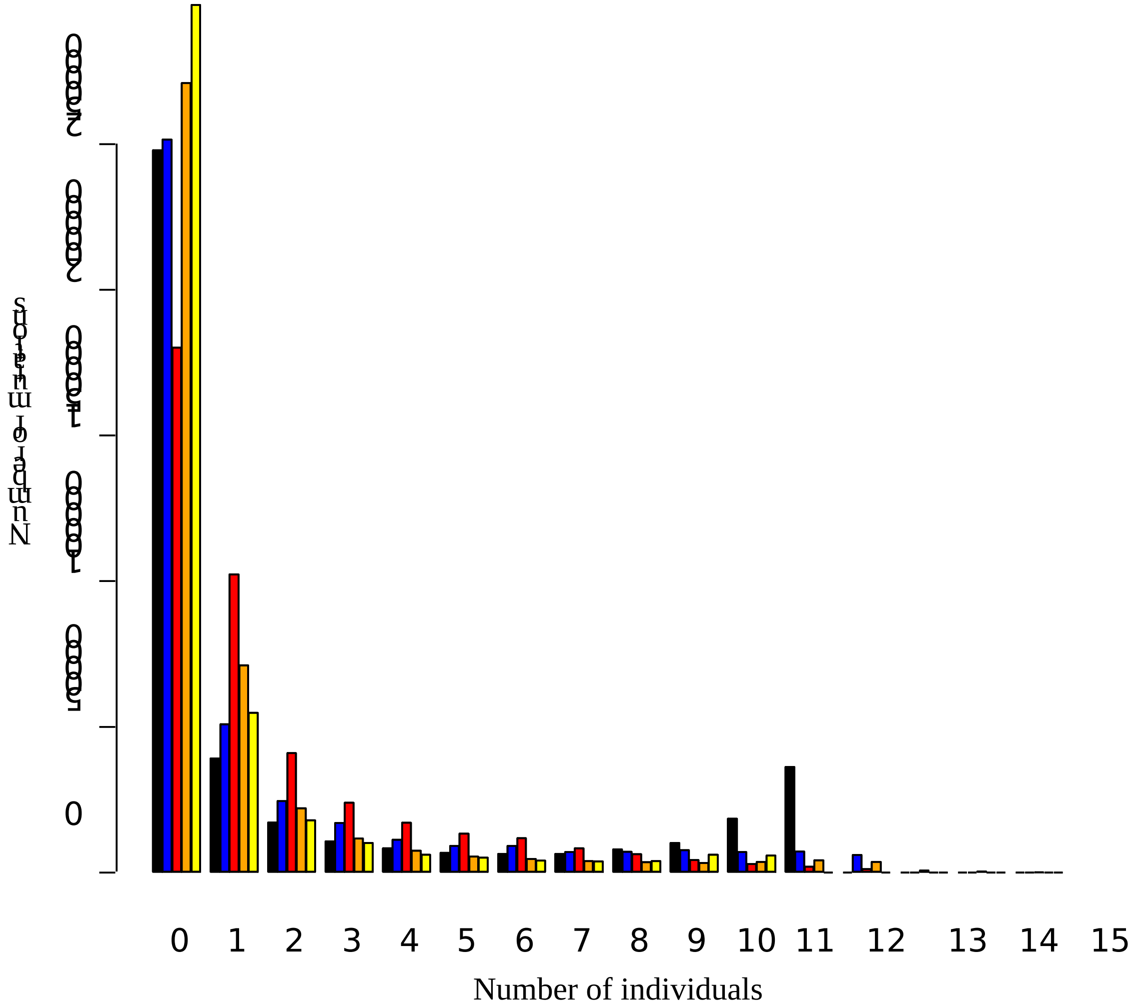


Figure S7. Folded site frequency spectra of single populations calculated including all minor alleles. Color code is the same as in Figure 1. Population Africa2 shows an excess of minor alleles in a high number of individuals compared to the other populations. The X-axis indicates the frequency of the minor allele. The Y-axis indicates the number of sites that fall in each frequency category per population.

Table S1. List of genomes analysed with geographic origin of the sample, name of the genome and number of base pairs covered by the sequencing. The genomes were assigned to populations America (Am), Asia (As), Africa1 (Af1), Africa2 (Af2), EurAsian (Eu) based on the DAPC analysis. The seven MLST housekeeping genes were assigned to ten of the previously published *H. pylori* (sub-)populations hpEastAsia (hspAmerind, hspEAsia), hpAsia2, hpEurope, hpNEAfrica, hpSahul, hpAfrica1 (hspWAfrica, hspSAfrica), and hpAfrica2 using Structure. The seven new genome sequences in this study are highlighted in bold.

Ind_N	Country	Genome Designation	Bases	DAPC (K=3)	BAPS (K=3)	STRUCTURE (K = 3) (100kb)	DAPC (K=4)	BAPS (K=4)	STRUCTURE (MLST)	Reference
1	Peru	Shi112	1,663,456	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	Kersulyte et al (2010)
2	Peru	Shi417	1,665,719	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
3	Peru	Shi169	1,616,909	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
4	Peru	Puno120	1,624,979	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
5	Peru	Puno135	1,646,139	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
6	Peru	PeCan18	1,660,685	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hpAfrica1	“
7	Peru	PeCan4	1,629,557	As-Am	As-Am	As-Am	As-Am	Eu-As	hpEurope	“
8	Peru	Cuz20	1,635,449	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
9	Peru	Shi470	1,608,548	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	“
10	Venezuela	v225d	1,588,278	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	Mane et al. (2010)
11	**Korea**	**DU15**	**1,614,411**	**As-Am**	**As-Am**	As-Am	**As-Am**	**As-Am**	**hspEAsia**	**Present study**
12	Japan	F16	1,575,399	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	Furuta et al. (2011)
13	Japan	F30	1,570,306	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
14	Japan	F32	1,578,824	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
15	Japan	F57	1,609,006	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
16	Korea	51	1,589,954	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	Kim et al. (2013)
17	Korea	52	1,568,826	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
18	Japan	OK113	1,616,617	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	Yahara et al. (2013)
19	Japan	OK310	1,591,278	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
20	Japan	35A	1,566,655	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
21	East Asia	83	1,617,426	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	“
22	China	XZ274	1,634,138	As-Am	As-Am	As-Am	As-Am	As-Am	hspEAsia	Guo et al. (2012)
23	India	India7	1,675,918	As-Am	As-Am	As-Am	Eu-As	Eu-As	hpAsia2	
24	India	SNT49	1,607,577	As-Am	As-Am	As-Am	Eu-As	Eu-As	hpAsia2	Kersulyte et al (2010)
25	**India**	**L7**	**1,617,826**	**As-Am**	**As-Am**	As-Am	**Eu-As**	**Eu-As**	**hpAsia2**	**Present study**

26	**Australian aboriginal**	**ausabrJ05**	**1,510,564**	**As-Am**	**As-Am**	As-Am	**Eu-As**	**Eu-As**	**hpSahul**	**Present study**
27	**Papua New Guinea**	**PNG84A**	**1,531,450**	**As-Am**	**As-Am**	As-Am	**Eu-As**	**Eu-As**	**hpSahul**	**Present study**
28	Gambia	Gambia94/24	1,709,911	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	Kersulyte et al (2010)
29	France (West African patient)	908	1,549,666	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	Devi et al (2010)
30	France (West African patient)	2017	1,548,238	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	Avasthi et a (2011)
31	France (West African patient)	2018	1,562,832	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	“
32	Peru	Sat464	1,560,342	As-Am	As-Am	As-Am	As-Am	As-Am	hspAmerind	Kersulyte et al (2010)
33	USA	J99	1,643,831	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	Alm et al (1999)
34	**Sudan**	**SU1**	**1,631,697**	**Af1-Eu**	**Af1-Eu**	Af1-Eu	**Eu-As**	**Eu-As**	**hpNEAfrica**	**Present study**
35	**South Africa**	**CC33C**	**1,659,899**	**Af1-Eu**	**Af1-Eu**	Af1-Eu	**Af1**	**Af1**	**hspSAfrica**	**Present study**
36	**South Africa (San)**	**K26A1**	**1,570,310**	**Af2**	**Af2**	Af2	**Af2**	**Af2**	**hpAfrica2**	**Present study**
37	South Africa	SouthAfrica7	1,653,913	Af2	Af2	Af2	Af2	Af2	hpAfrica2	Kersulyte et al (2010)
38	Lithuania	Lithuania75	1,624,644	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	“
39	Italy	G27	1,652,982	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	Baltrus et al. (2009)
40	U.S.	B8	1,673,997	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	Farnbacher et al (2010)
41	Europe	P12	1,673,813	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	Fischer et al. 2008
42	Sweden	HPAG1	1,596,366	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	Oh et al.

										2006
43	UK	26695	1,667,867	Af1-Eu	Af1-Eu	As-Am	Eu-As	Eu-As	hpEurope	Tomb et al. 1997
44	Spain	HUP-B14	1,599,280	Af1-Eu	Af1-Eu	Af1-Eu	Eu-As	Eu-As	hpEurope	Kersulyte et al (2012)
45	El Salvador	ELS37	1,664,587	Af1-Eu	Af1-Eu	Af1-Eu	Eu-As	Eu-As	hpEurope	
46	Peru	SJM180	1,658,051	Af1-Eu	Af1-Eu	Af1-Eu	Eu-As	Eu-As	hpEurope	Kersulyte et al (2010)
47	South Africa	SA29A	1,620,792	Af2	Af2	Af2	Af2	Af2	hpAfrica2	Didelot et al. (2013)
48	South Africa	SA29C	1,650,643	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	“
49	South Africa	SA30A	1,670,386	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	**hspSAfrica**	“
50	South Africa	SA47A	1,670,767	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
51	South Africa	SA144A	1,630,166	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
52	South Africa	SA155A	1,664,256	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
53	South Africa	SA160A	1,658,851	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
54	South Africa	SA161C	1,689,135	Af1	Af1	Af1-Eu	Af1	Af1	hspWAfrica	“
55	South Africa	SA166A	1,599,095	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
56	South Africa	SA169A	1,652,206	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
57	South Africa	SA172C	1,658,069	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
58	South Africa	SA194A	1,565,205	Af2	Af2	Af2	Af2	Af2	hpAfrica2	“
59	South Africa	SA224A	1,645,150	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	hspWAfrica	“
60	South Africa	SA226A	1,705,370	Af1-Eu	Af1-Eu	Af1-Eu	Af1	Af1	**hspSAfrica**	“

Table S2A. Maximum likelihood model selection among five scenarios describing the global demographic history of *Helicobacter pylori. K* is the number of model parameters; AICc is the corrected Akaike information criterion; Δi is the difference between the AICc calculated for a specific model and the minimum AICc calculated among all models. For results on all migration models tested see STable 2.

N POP	Model	Topology	Parallel runs	Best log likelihood	K	AICc	Δi
Three pops	Divergence	Tree topology 1	100	-2043685	12	9413237.11	9413240.69
“	“	Tree topology 2	100	-2016635	12	**9288644.81**	**9288648.39**
“	“	Tree topology 3	100	-2017323	12	9291813.73	9291817.32
Four pops	“	Tree topology 1	100	-3660034	16	16858140.60	16858144.19
“	“	Tree topology 2	100	-3650047	16	**16812148.48**	**16812155.03**
“	“	Tree topology 3	100	-3679940	16	16949835.64	16949842.19
Five pops	“	Tree topology 1	100	-5336876	20	24581682.85	24581689.41
“	“	Tree topology 2	100	-5326078	20	**24531947.26**	**24531953.82**
“	“	Tree topology 3	100	-5338409	20	24588751.85	24588762.48
“	Migration	Tree topology 2	67	-5178895	40	**23854070.37**	**23854196.52**

Table S2B. Demographic models with migration tested and relatives likelihood results. K = number of parameters, Max log10 lhood = best likelihood over N independent runs, AIC = Akaike Information Criterion and corrected AIC.

Model	K	Max log10 lhood	N	AIC	AICc	Description
4 pops SS + 3 pops IS (Af1, Eu, As)	32	-5264244	69	24247171.864	24247230.531	#mig between all paired (-Af2) pops plus 3 interpaired
4 pops SS + 2 pops IS (Af1, Eu)	30	-5266370	80	24256960.22	24256998.179	#mig between all paired (-Af2) pops plus 2 interpaired
4 pops SS + 2 pops IS (Af1, As)	28	-5272008	60	24282924.848	24282977.235	#mig between all paired (-Af2) pops plus Af-As
4 pops SS (Af2, Af1, Eu, As)	24	-5270720	66	24276984.32	24277013.588	#mig between paired pops Af2-As
5 pops SS (Af2, Af1, Eu, As, Am)	26	-5274989	46	24296651.334	24296725.228	#mig between paired pops Af2-Am

5 pops SS (Af2, Af1, Eu, As, Am)	26	-5268775	60	24268029.65	24268072.195	#mig between paired pops Af2-Am plus Af2-As
5 pops SS + 3 pops IS (Af2, Af1, Eu)	30	-5195755	24	23931707.53	23931441.816	#mig between all paired pops plus Af2-Eu
5 pops SS + 4 pops IS (Af2, Af1, Eu, As)	32	-5195226	34	23929274.956	23931386.956	#mig between all paired pops plus Af2-Eu Af1-As
5 pops SS + 3 pops IS (Af2, Af1, Eu)	34	-5189880	75	23904655.28	23904714.78	#mig between all paired pops plus Af2-Eu Af1-As, Eu-Am
5 pops SS + 3 pops IS (Af2, Af1, Eu, As)	36	-5186701	100	23890016.806	23890059.092	#mig between all paired pops plus Af2-Eu, Af1-As, Eu-Am, Af2-As
5 pops IS but pair Af2-Am	38	-5181030	50	23863900.18	23864169.635	#mig between all paired pops plus all interpaired (- (Af2-Am))
5 pops IS	40	-5178895	67	23854070.37	23854196.524	#mig between all paired pops plus all interpaired

SS = stepping stone
IS = island
AIC = 2k -2ln(likelihood)
AICc = AIC + 2k(k+1)/(n-k-1)

Table S3. Main genes under selection by population.

Gene	***Position***	***Protein***	***Populations***					
			Wd	Af2	Af1	EuAs	EaAs	Am
mtnN	65475-64750	MTA/SAH nucleosidase	*					
Unknown	75528-76634	Uncharacterized protein HI_0882	*					
msrAB	186603-187625	Peptide methionine sulfoxide reductase MsrA/MsrB	*					
typA	397037-398857	GTP-binding protein TypA/BipA homolog	*					*
bsp6IM	400156-400575	Modification methylase Bsp6I	*		*			*
ykpA	426574-428175	Uncharacterized ABC transporter ATP-binding protein YkpA	*					*
coaBC	431330-430053	Coenzyme A biosynthesis bifunctional protein CoaBC	*					
dus	532116-533102	Probable tRNA-dihydrouridine synthase	*				*	
rpoN	540290-541495	RNA polymerase sigma-54 factor	*				*	
rsmH	544664-543756	Ribosomal RNA small subunit methyltransferase H	*				*	
aroC	583608-582511	Chorismate synthase	*					
rnhA	584745-584281	Ribonuclease H	*					
fliY	668666-669037	Flagellar motor switch phosphatase FliY	*					
yebA	711089-709860	Uncharacterized metalloprotease HI_0409	*			*		
yfbE	870116-871246	Uncharacterized protein MJ1066	*					
valS	925424-922806	Valyl-tRNA synthetase	*					
copA	1188896-1186560	Copper-exporting P-type ATPase A	*	*	*	*		
nusA	1200135-1201322	Transcription elongation protein nusA	*			*		
mod	1202292-1201546	Type III restriction-modification system EcoP15I enzyme mod	*			*	*	
sdaA	104731-103364	L-serine dehydratase		*				
sdaC	105918-104731	Serine transporter		*				
hemA	202001-200658	Glutamyl-tRNA reductase		*				
pyrB	310635-311558	Aspartate carbamoyltransferase		*				
Unknown	401880-403466	Outer Membrane Protein		*				
hup	436148-435789	DNA-binding protein HU		*				
Unknown	545671-544850	Outer Membrane Protein		*				

korC	661401-660841	2-oxoglutarate synthase subunit korC	*		*	*
fabF	689739-690980	3-oxoacyl-[acyl-carrier-protein] synthase 2	*			
accA	691000-691956	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha	*			
Unknown	764984-765604	Cytotoxin-Protein like vacA	*			
tdcF	784391-784014	RutC family protein jhp_0879	*			
Unknown	788720-787707	Conserved Hypothetical Protein	*			
ftsA	812431-813924	Cell division protein ftsA	*			
ngoBIM	901339-900317	Modification methylase NgoBI	*			
fmt	916259-915348	Methionyl-tRNA formyltransferase	*		*	
deoD	953627-952896	Purine nucleoside phosphorylase deoD-type	*			
fliI	1120192-1118888	Flagellum-specific ATP synthase	*			
Unknown	1124260-1124628	Hypothetical Protein HPSJM	*			
Unknown	1128984-1126906	Ribonuclease J	*			
mnmE	1136961-1138346	tRNA modification GTPase MnmE	*			
ybhG	1173501-1172512	36 kDa antigen	*			
bdlA	58938-56917	Biofilm dispersion protein BdlA		*		
serA	70004-70948	Putative 2-hydroxyacid dehydrogenase HI_1556		*		
frdA	156967-154823	Fumarate reductase flavoprotein subunit		*		
pyrG	291361-289832	CTP synthase		*		
msbA	313338-314993	Lipid A export ATP-binding/permease protein MsbA		*	*	
nixA	316216-317211	High-affinity nickel-transport protein nixA		*		
copA	321905-319935	Copper-transporting ATPase		*	*	
Unknown	364539-363703	Hypothetical Protein Jhp		*		
tdhA	408440-406221	TonB-dependent heme receptor A		*		
pgbB	418961-420319	Plasminogen-binding protein pgbB		*		
flhB	496895-495987	Flagellar biosynthetic protein flhB		*		
acxA	559085-556944	Acetone carboxylase beta subunit		*		
vacA	731291-732442	Vacuolating cytotoxin autotransporter		*	*	
bioA	810846-809503	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase		*		
torZ	830889-830026	Trimethylamine-N-oxide reductase		*		*

edd	880675-878843	Phosphogluconate dehydratase	*			
rplS	919545-918160	50S ribosomal protein L19	*	*		
yhhG	1086396-1085875	Putative nickel-responsive regulator	*			
rlmN	1125936-1124869	Ribosomal RNA large subunit methyltransferase N	*			
Unknown	1138544-1140799	Outer membrane protein	*			
yrbH	1126922-1125933	Uncharacterized protein HP_1429	*			
mboIR	67299-67616	Type-2 restriction enzyme MboI		*		
mboIBM	67666-68421	Modification methylase MboIB		*	*	
yebC	131914-131192	UPF0082 protein HPG27_148		*		
yqhA	152030-152563	UPF0114 protein HPSH_00970		*		
rfaJ	170222-169350	Lipopolysaccharide Biosynthesis Protein		*		
vacA	247650-249185	Vacuolating cytotoxin autotransporter		*		
copP	319934-319734	COP-associated protein		*		
cadA	471626-473686	Cadmium zinc and cobalt-transporting ATPase		*		
pgpA	524100-523648	Phosphatidylglycerophosphatase A		*		
msbA2	651289-651711	Lipid A export ATP-binding/permease protein MsbA 2		*		
dcrA	651783-653084	Chemoreceptor protein A		*		
bioF	653104-654228	Putative 8-amino-7-oxononanoate synthase/2-amino-3-ketobutyrate coenzyme A ligase		*		
mrcA	654228-656207	Penicillin-binding protein 1A		*		
Unknown	656788-656210	Tumor Necrosis Factor Alpha-Inducing Protein		*		*
gyrB	714297-711976	DNA gyrase subunit B		*		
dnaN	715432-714308	DNA polymerase III subunit beta		*		
soj	914716-913916	Sporulation initiation inhibitor protein soj		*		
birA	915351-914713	Biotin-Protein Ligase		*		
trmD	920231-919542	tRNA guanine-N1--methyltransferase		*		
pgi	938212-939876	Glucose-6-phosphate isomerase		*		
Unknown	951951-950851	Outer Membrane Protein		*		
tal	1180463-1181413	Transaldolase		*		
Unknown	1185912-1184782	Outer Membrane Protein		*	*	*
yqiI	493880-495163	Uncharacterized protein YqiI			*	

tilS	531009-532025	tRNAIle-lysidine synthase	*	
yhdH	719232-717904	Uncharacterized sodium-dependent transporter yhdH	*	
dps	204062-203625	DNA protection during starvation protein		*
flgL	257020-254537	Flagellar hook-associated protein 3		*
murC	620217-621566	UDP-N-acetylmuramate--L-alanine ligase		*
yecS	779424-778525	Probable amino-acid ABC transporter permease protein HI_0179		*
accC	868087-869406	Biotin carboxylase		*

Table S4. Selected genes in Worldwide and local populations with their position along the chromosome and the selection value associated to the region. Omega refers to the linkage disequilibrium value and Alpha to the likelihood value indicated by SweeD. Most relevant genes are highlighted in bold.

Worldwide						
Selection value		**Selected region**				
Omega	*Alpha*	from	to	*Genes*	*Position*	*Protein*
5.007-7.206		64534	66142	*rpoD*	64726-63008	RNA polymerase sigma factor rpoD
				mtnN	65475-64750	MTA/SAH nucleosidase
				fabD	66415-65486	Malonyl CoA-acyl carrier protein transacylase
5.029-5.776		76042	76192	***Unknown***	75528-76634	Uncharacterized protein HI_0882
5.015-12.292		185916	189132	*radA*	185077-186423	DNA repair protein RadA homolog
				msrAB	186603-187625	Peptide methionine sulfoxide reductase MsrA/MsrB
5.010-9.698		396753	402035	*vspIM*	394565-397009	Modification methylase VspI
				typA	397037-398857	GTP-binding protein TypA/BipA homolog
				bsp6IM	400156-400575	Modification methylase Bsp6I
				srpA	401653-400709	Protein SrpA
5.000-25.206		423991	437615	*hldE*	423432-424817	Bifunctional protein hldE
				gmhA	424810-425388	Phosphoheptose isomerase
				guaC	426385-425402	GMP reductase
				ykpA	426574-428175	Uncharacterized ABC transporter ATP-binding protein YkpA
				coaBC	431330-430053	Coenzyme A biosynthesis bifunctional protein CoaBC
5.006-10.061		532487	538473	***dus***	532116-533102	Probable tRNA-dihydrouridine synthase
5.011-10.609		540310	544098	***rpoN***	540290-541495	RNA polymerase sigma-54 factor
				rsmH	544664-543756	Ribosomal RNA small subunit methyltransferase H
5.016-5.373		582320	585510	***aroC***	583608-582511	Chorismate synthase
				rnc	584222-583605	Ribonuclease 3
				rnhA	584745-584281	Ribonuclease H
5.012-12.900		663746	665899	*yrrL*	664861-663866	UPF0755 protein yrrL
5.002 -13-383		667457	669814	*nth*	667943-668584	Endonuclease III

				fliY	668666-669037	Flagellar motor switch phosphatase FliY
5.001-5.439		710394	710632	***yebA***	711089-709860	Uncharacterized metalloprotease HI_0409
5.003-5.156		870594	870822	***yfbE***	870116-871246	Uncharacterized protein MJ1066
5.002-5.880		922656	924350	*ffh*	922792-921446	Signal recognition particle protein
				valS	925424-922806	Valyl-tRNA synthetase
5.003-9.112		1183195	1185380	***copA***	1188896-1186560	Copper-exporting P-type ATPase A
6.483-42.134		1201740	1203910	***nusA***	1200135-1201322	Transcription elongation protein nusA
				mod	1202292-1201546	Type III restriction-modification system StyLTI enzyme mod
				recG	1204193-1202334	ATP-dependent DNA helicase recG
						Africa2
Selection value		**Selected region**				
Omega	*Alpha*	from	to	*Genes*	*Position*	*Protein*
	2.211-7.934	30133	30141	*Unknown*	30198-30569	Component Of Conjugal Plasmid Transfer System
2.581-6.079	1.891-8.612	103022	107749	***sdaA***	104731-103364	L-serine dehydratase
				sdaC	105918-104731	Serine transporter
				aroH	106150-107499	Phospho-2-dehydro-3-deoxyheptonate aldolase
	1.891-35.162	201751	201802	***hemA***	202001-200658	Glutamyl-tRNA reductase
2.330-3.860	1.891-143.81	311466	311878	***pyrB***	310635-311558	Aspartate carbamoyltransferase
5.545-9.578	2.145-2.145	401610	403641	***Unknown***	401880-403466	Outer Membrane Protein
2.454-5.979		435390	435724	***hup***	436148-435789	DNA-binding protein HU
5.608-5.777		545415	545446	***Unknown***	545671-544850	Outer Membrane Protein
2-376-2.393	1.306-104.02	660253	660444	***korC***	661401-660841	2-oxoglutarate synthase subunit korC
	1.908-16.730	689670	691997	***fabF***	689739-690980	3-oxoacyl-[acyl-carrier-protein] synthase 2
2.209-3.077	5.815-48.923	691926	691997	***accA***	691000-691956	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha
	1.891-51.600	764639	765817	***Unknown***	764984-765604	Cytotoxin-Protein like vacA
	1.891-87.835	784384	784539	***tdcF***	784391-784014	RutC family protein jhp_0879
	10.23-219.00	787341	787761	***Unknown***	788720-787707	Conserved Hypothetical Protein
	6.401-151.31	813199	813453	***ftsA***	812431-813924	Cell division protein ftsA
	1.891-18.398	900344	900397	**ngoBIM**	901339-900317	Modification methylase NgoBI

5.567-5.567		916255	916270	***fmt***	916259-915348	Methionyl-tRNA formyltransferase
	1.306-30.109	952216	952930	***deoD***	953627-952896	Purine nucleoside phosphorylase deoD-type
		959795	959959	*yjbQ*	959689-958538	Putative Na+/H+ antiporter yjbQ
	1.302-14-249	1119163	1119392	***fliI***	1120192-1118888	Flagellum-specific ATP synthase
	2.818-9.705	1124230	1124250	***Unknown***	1124260-1124628	Hypothetical Protein HPSJM
	3.096-7.884	1128807	128822	***Unknown***	1128984-1126906	Ribonuclease J
	7.663-30.008	1139567	1139671	***mnmE***	1136961-1138346	tRNA modification GTPase MnmE
				lpp20	1142347-1141820	LPP20 lipoprotein
				trxA	1143664-1143350	Thioredoxin
				yciL	1144484-1143984	Uncharacterized RNA pseudouridine synthase HP_1459
				dnaE	1148101-1144466	DNA polymerase III subunit alpha
2.208-6.017	1.506-86.617	1173804	1173982	***ybhG***	1173501-1172512	36 kDa antigen
		1183130	1188327	***copA***	1188896-1186560	Copper-exporting P-type ATPase A
	1.891-25.489	1201513	1207760	***mod***	1202292-1201531	Type III restriction-modification system EcoP15I enzyme mod

Africa1

Selection value		**Selected region**				
Omega	*Alpha*	from	to	*Genes*	*Position*	*Protein*
	1.949-4.368	55702	57253	***bdlA***	58938-56917	Biofilm dispersion protein BdlA
2.314-2.347	2.030-20.716	70957	70991	***serA***	70004-70948	Putative 2-hydroxyacid dehydrogenase HI_1556
	1.950-17.106	156171	156204	***frdA***	156967-154823	Fumarate reductase flavoprotein subunit
	1.951-15.519	291514	291654	***pyrG***	291361-289832	CTP synthase
	1.951-6.605	311772	317338	***nixA***	316216-317211	High-affinity nickel-transport protein nixA
3.127-6.074	1.950-1.950	318397	320617	***copA***	321905-319935	Copper-transporting ATPase
	1.950-17.109	363917	364087	***Unknown***	364539-363703	Hypothetical Protein Jhp
	1.950-1.950	378546	381376	*Unknown*	381536-378168	Putative type-1 restriction enzyme MjaXP R protein
	4.013-10.477	400638	400653	***bsp6IM***	400156-400575	Modification methylase Bsp6I
	1.950-12.856	408624	408891	***tdhA***	408440-406221	TonB-dependent heme receptor A
				katA	408905-410422	Catalase
	2.169-2.169	420286	420520	***pgbB***	418961-420319	Plasminogen-binding protein pgbB

	1.950-7.605	495165	496612	***flhB***	496895-495987	Flagellar biosynthetic protein flhB
	1.950-61.061	559094	560091	*Unknown*	560264-559263	Outer Membrane Protein
				acxA	559085-556944	Acetone carboxylase beta subunit
	1.950-235.76	731666	737913	***vacA***	731291-732442	Vacuolating cytotoxin autotransporter
	2.257-18.600	810425	810891	***bioA***	810846-809503	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
	1.950-7.429	829916	829937	***torZ***	830889-830026	Trimethylamine-N-oxide reductase
	2.253-6.678	878842	878849	***edd***	880675-878843	Phosphogluconate dehydratase
4.616-6.167		918133	919152	***rplS***	919545-918160	50S ribosomal protein L19
	1.950-10.205	1086358	1086373	***yhhG***	1086396-1085875	Putative nickel-responsive regulator
	1.950-17.598	1124806	1124831	***rlmN***	1125936-1124869	Ribosomal RNA large subunit methyltransferase N
	1.950-6.636	1138604	1140513	***Unknown***	1138544-1140799	Outer membrane protein
2.981-3.398	1.950-2.209	1184394	1187707	***copA***	1188908-1186560	Copper-exporting P-type ATPase A
	2.214-8.728	1265134	1265144	***yrbH***	1126922-1125933	Uncharacterized protein HP_1429

EuroAsian

Selection value		**Selected region**				
Omega	*Alpha*	from	to	*Genes*	*Positon*	*Protein*
3.016-5.534	3.205-3.427	65560	68628	*fabD*	66415-65486	Malonyl CoA-acyl carrier protein transacylase
				mboIR	67299-67616	Type-2 restriction enzyme MboI
				mboIBM	67666-68421	Modification methylase MboIB
2.015-2.522		130716	132401	***yebC***	131914-131192	UPF0082 protein HPG27_148
2.004-2.577		150257	153486	***yqhA***	152030-152563	UPF0114 protein HPSH_00970
				ymdC	153879-152569	Uncharacterized protein jhp_0176
2.000-3.160		164132	171689	*mrp*	168297-169403	Protein mrp homolog
				rfaJ	170222-169350	Lipopolysaccharide Biosynthesis Protein
2.000-2.861		245882	249201	***vacA***	247650-249185	Vacuolating cytotoxin autotransporter
2.002-2.421		318051	320366	***copP***	319934-319734	COP-associated protein
				copA	321905-319935	Copper-transporting ATPase
2.004-2.469		472673	474939	***cadA***	471626-473686	Cadmium zinc and cobalt-transporting ATPase
2.004-3.413		523543	527529	***pgpA***	524100-523648	Phosphatidylglycerophosphatase A

				rimO	525847-527163	Ribosomal protein S12 methylthiotransferase RimO
2.000-5.480		651419	664090	***msbA2***	651289-651711	Lipid A export ATP-binding/permease protein MsbA 2
				dcrA	651783-653084	Chemoreceptor protein A
				bioF	653104-654228	Putative 8-amino-7-oxononanoate synthase/2-amino-3-ketobutyrate coenzyme A ligase
				mrcA	654228-656207	Penicillin-binding protein 1A
				Unknown	656788-656210	Tumor Necrosis Factor Alpha-Inducing Protein
				korC	661401-658738	2-oxoglutarate synthase subunit korC
2.001-4.382		708670	715886	*nudK*	709860-709222	GDP-mannose pyrophosphatase nudK
				yebA	711089-709860	Uncharacterized metalloprotease HI_0409
				gyrB	714297-711976	DNA gyrase subunit B
				dnaN	715432-714308	DNA polymerase III subunit beta
				pldA	715856-715488	Phospholipase A
2.000-2.353		730069	733616	***vacA***	730306-731331	Vacuolating cytotoxin autotransporter
3.002-4.783	1.913-2.939	911000	922492	*atpA*	911417-909906	ATP synthase subunit alpha
				atpF	912498-911983	ATP synthase subunit b
				parB	913916-913041	Probable chromosome-partitioning protein parB
				soj	914716-913916	Sporulation initiation inhibitor protein soj
				birA	915351-914713	Biotin-Protein Ligase
				fmt	916259-915348	Methionyl-tRNA formyltransferase
				rplS	919545-918160	50S ribosomal protein L19
				trmD	920231-919542	tRNA guanine-N1--methyltransferase
2.001-2.564		939548	940261	***pgi***	938212-939876	Glucose-6-phosphate isomerase
2.000-3.109		950913	952986	***Unknown***	951951-950851	Outer Membrane Protein
2.899-3.555		1181098	1181803	***tal***	1180463-1181413	Transaldolase
2.003-6.186		1182553	1187171	*Unknown*	1182582-1183649	Permease YjgP/YjgQ
				Unknown	1185912-1184782	Outer Membrane Protein
				copA	1188896-1186560	Copper-exporting P-type ATPase A
2.000-2.434		1200508	1202493	***nusA***	1200135-1201322	Transcription elongation protein nusA

EastAsia

Selected value		**Selected Region**				
Omega	*Alpha*	From	To	*Genes*	*Position*	*Protein*
2.002-4.008		66774	69397	***mboIBM***	67666-68421	Modification methylase MboIB
3.465-5.527	3.464-5.527	495168	495173	***yqiI***	493880-495163	Uncharacterized protein YqiI
4.061-5.244		530852	534109	***tilS***	531009-532025	tRNAIle-lysidine synthase
				dus	532116-533102	Probable tRNA-dihydrouridine synthase
6.634-11.768		540372	544044	***rpoN***	540290-541495	RNA polymerase sigma-54 factor
				rsmH	544601-543756	Ribosomal RNA small subunit methyltransferase H
6.621-8.421		718425	718827	***yhdH***	719232-717904	Uncharacterized sodium-dependent transporter yhdH
5.702-6.608		1184465	1184778	***Unknown***	1185912-1184782	Outer Membrane Protein
1.958-3.982		1201505	1202312	***mod***	1202292-1201531	Type III restriction-modification system EcoP15I enzyme mod

America

Selected value		**Selected Region**				
Omega	*Alpha*	From	To	*Genes*	*Position*	*Protein*
	1.478-13.226	204197	204232	***dps***	204062-203625	DNA protection during starvation protein
				atoS	205346-204240	Signal transduction histidine-protein kinase AtoS
	2.546-10.332	257094	257117	***flgL***	257020-254537	Flagellar hook-associated protein 3
1.909-2.033	1.478-209.729	398847	400627	***typA***	397037-398857	GTP-binding protein TypA/BipA homolog
				bsp6IM	400156-400575	Modification methylase Bsp6I
3.227-11.493	1.478-1.478	423963	427398	*hldE*	423432-424817	Bifunctional protein hldE
				gmhA	424810-425388	Phosphoheptose isomerase guaC GMP reductase
				ykpA	426574-428175	Uncharacterized ABC transporter ATP-binding protein YkpA
	5.289-5.557	621922	621929	***murC***	620217-621566	UDP-N-acetylmuramate--L-alanine ligase
3.125-4.223	1.478-1.478	656156	656870	***Unknown***	656788-656210	Tumor Necrosis Factor Alpha-Inducing Protein
6.356-6.603		661347	661631	***korC***	661401-658738	2-oxoglutarate synthase subunit korC
	1.478-25.209	778456	778560	***yecS***	779424-778525	Probable amino-acid ABC transporter permease protein HI_0179
				Unknown	786427-786744	Na+/H+ Antiporter NhaC
	0.977-4.213	826963	827825	*ynaI*	826037-824169	Uncharacterized mscS family protein jhp_0969

				guaA	829409-827883	GMP synthase [glutamine-hydrolyzing]
	1.719-5.545	830017	832424	***torZ***	830973-830026	Trimethylamine-N-oxide reductase
	1.478-7.242	869409	869429	***accC***	868087-869406	Biotin carboxylase
1.984-6.338		1183649	1184680	*Unknown*	1182582-1183649	Permease YjgP/YjgQ
				Unknown	1185912-1184782	Outer Membrane Protein

File S1

Fastsimcoal2.1 input file est

```
// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name   #dist.#min  #max
//all Ns are in number of haploid individuals
1 NA0 unif 10000 500000 output
1 NA1 unif 10000 500000 output
1 NA2 unif 10000 500000 output
1 NA3 unif 10000 500000 output
1 NA4 unif 10000 500000 output
1 NPOP0 unif 100000 1000000 output
1 NPOP1 unif 100000 1000000 output
1 NPOP2 unif 100000 1000000 output
1 NPOP3 unif 100000 1000000 output
1 NPOP4 unif 100000 1000000 output
1 DIV4 unif 10000 30000 output
1 DIV3 unif 30000 50000 output
1 DIV2 unif 50000 80000 output
1 DIV1 unif 80000 250000 output
0 MUT unif 0.0000001 0.001 output
0 M01 unif 0.0000001 0.00001 output
0 M10 unif 0.0000001 0.00001 output
0 M12 unif 0.0000001 0.00001 output
0 M21 unif 0.0000001 0.00001 output
0 M23 unif 0.0000001 0.00001 output
0 M32 unif 0.0000001 0.00001 output
0 M34 unif 0.0000001 0.00001 output
0 M43 unif 0.0000001 0.00001 output
0 MASAF unif 0.0000001 0.00001 output
0 MAFAS unif 0.0000001 0.00001 output
0 MEUAF unif 0.0000001 0.00001 output
```

```
0 MAFEU unif 0.0000001 0.00001 output
0 MEUAM unif 0.0000001 0.00001 output
0 MAMEU unif 0.0000001 0.00001 output
0 AF2AS unif 0.0000001 0.00001 output
0 ASAF2 unif 0.0000001 0.00001 output
0 AMAF1 unif 0.0000001 0.00001 output
0 AF1AM unif 0.0000001 0.00001 output
0 LAST unif 0.0000001 0.00001 output
0 TSAL unif 0.0000001 0.00001 output

[RULES]

[COMPLEX PARAMETERS]
0 T0 = NA0/NPOP0 hide
0 T1 = NA1/NPOP1 hide
0 T2 = NA2/NPOP2 hide
0 T3 = NA3/NPOP3 hide
0 T4 = NA4/NPOP4 hide
0 L0 = log(T0) hide
0 L1 = log(T1) hide
0 L2 = log(T2) hide
0 L3 = log(T3) hide
0 L4 = log(T4) hide
0 R0 = L0/DIV1 output
0 R1 = L1/DIV1 output
0 R2 = L2/DIV2 output
0 R3 = L3/DIV3 output
0 R4 = L4/DIV4 output
```

File S2
Fastsimcoal2.1 input file tpl

```
//Number of populations
5
//Deme sizes (haploid number of genes)
NPOP0
NPOP1
NPOP2
NPOP3
NPOP4
//Sample sizes
10
10
16
12
9
//Growth rates
R0
R1
R2
R3
R4
//Number of migration matrices : If 0 : No migration between demes
5
//Migration matrix 0
0.0000 M01 MAFEU AF2AS TSAL
M10 0.0000 M12 MAFAS AF1AM
MEUAF M21 0.0000 M32 MEUAM
ASAF2 MASAF M23 0.0000 M43
LAST AMAF1 MAMEU M34 0.0000
//Migration matrix 1
0.0000 M01 MAFEU AF2AS 0.0000
M10 0.0000 M12 MAFAS 0.0000
MEUAF M21 0.0000 M32 0.0000
```

```
ASAF2 MASAF M23 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
//Migration matrix 2
0.0000 M01 MAFEU 0.0000 0.0000
M10 0.0000 M12 0.0000 0.0000
MEUAF M21 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
//Migration matrix 3
0.0000 M01 0.0000 0.0000 0.0000
M10 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
//Migration matrix 4
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
0.0000 0.0000 0.0000 0.0000 0.0000
//Historical event: time, source, sink, proportion of migrants, new deme size, new growth rate, new migration matrix
4 historical event
DIV4 4 3 1 1 0 1
DIV3 3 2 1 1 0 2
DIV2 2 1 1 1 0 3
DIV1 1 0 1 1 0 4
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of linkage blocks
1
//per Block: data type, num loci, rec. rate and mut rate + optional parameters
FREQ 1 0 MUT OUTEXP
```

Files S3-S7

Available for download as .txt files at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176404/-/DC1

File S3 Africa 1 Gene annotation

Fie S4 Africa 2 Gene annotation

File S5 Europe Gene annotation

File S6 Asia Gene annotation

File S7 America Gene annotation