

# An African origin for the intimate association between humans and *Helicobacter pylori*

Bodo Linz<sup>1\*</sup>, François Balloux<sup>2\*</sup>, Yoshan Moodley<sup>1</sup>, Andrea Manica<sup>3</sup>, Hua Liu<sup>2</sup>, Philippe Roumagnac<sup>1</sup>, Daniel Falush<sup>4</sup>, Christiana Stamer<sup>1</sup>, Franck Prugnolle<sup>5</sup>, Schalk W. van der Merwe<sup>6</sup>, Yoshio Yamaoka<sup>7</sup>, David Y. Graham<sup>7</sup>, Emilio Perez-Trallero<sup>8</sup>, Torkel Wadstrom<sup>9</sup>, Sebastian Suerbaum<sup>10</sup> & Mark Achtman<sup>1</sup>

Infection of the stomach by *Helicobacter pylori* is ubiquitous among humans. However, although *H. pylori* strains from different geographic areas are associated with clear phylogeographic differentiation<sup>1–4</sup>, the age of an association between these bacteria with humans remains highly controversial<sup>5,6</sup>. Here we show, using sequences from a large data set of bacterial strains that, as in humans, genetic diversity in *H. pylori* decreases with geographic distance from east Africa, the cradle of modern humans. We also observe similar clines of genetic isolation by distance (IBD) for both *H. pylori* and its human host at a worldwide scale. Like humans, simulations indicate that *H. pylori* seems to have spread from east Africa around 58,000 yr ago. Even at more restricted geographic scales, where IBD tends to become blurred, principal component clines in *H. pylori* from Europe strongly resemble the classical clines for Europeans described by Cavalli-Sforza and colleagues<sup>7</sup>. Taken together, our results establish that anatomically modern humans were already infected by *H. pylori* before their migrations from Africa and demonstrate that *H. pylori* has remained intimately associated with their human host populations ever since.

Over half of all humans are infected by *Helicobacter pylori*, a Gram-negative bacterium that can cause peptic ulcers and constitutes a risk factor for stomach cancer<sup>8</sup>. Not only is *H. pylori* ubiquitous, but it also possesses strong phylogeographic structure<sup>1</sup>, suggesting that bacterial polymorphisms reflect human phylogeography and historical migrations<sup>2,3,5</sup>. In 2003, we assigned 370 *H. pylori* strains to four main population clusters, two of which were subdivided into subpopulations<sup>2</sup>. The geographic sources of these strains reflected major events in human settlement history, such as the colonisation of Polynesia and the Americas and the African Bantu migrations. However, these discrete groupings seem to contradict an apparent continuity of the geographic component of genetic diversity in humans: genetic differentiation between human populations increases linearly with geographic distance computed along landmasses<sup>9–12</sup>; and their genetic diversity declines with increasing geographic distance from east Africa<sup>13,14</sup>.

There are several possible explanations why detailed population genetic patterns differ between *H. pylori* and their human hosts. Infection of humans by *H. pylori* might be too recent to have been affected by ancient events in human settlement history, for example, it might date from a recently acquired zoonosis<sup>5</sup>. Or differences in

population structure between bacteria and humans may reflect frequent horizontal transmission of *H. pylori*. Alternatively, apparent differences in the population genetic patterns may simply be a matter of perception owing to differing analytical methodology: *H. pylori* population genetics has so far focused on the description of clusters, whereas human population genetics is influenced by a traditional emphasis on clines<sup>15</sup>.

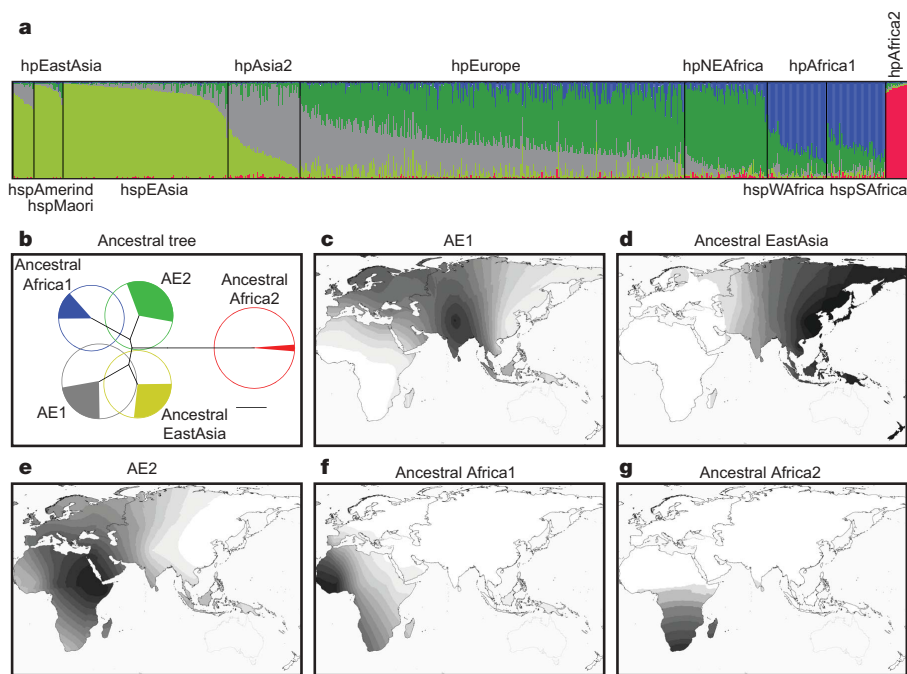
We used an expanded data set (769 *H. pylori* isolates from 51 ethnic sources; Supplementary Table 1) to test whether patterns in their geographic distribution mimic those of humans. Bayesian MCMC (Monte-Carlo Markov chain) cluster analyses identified the same five ancestral sources of nucleotides as found previously with a smaller data set<sup>2</sup> (Fig. 1a, b). These analyses also assigned the isolates to six populations containing various degrees of ancestry from the five ancestral sources (Fig. 1a; Supplementary Fig. 1a). Four of the populations had been previously identified, and designated hpEurope, hpEastAsia, hpAfrica1 and hpAfrica2 owing to their obvious geographical associations<sup>2</sup>. In agreement, almost all strains isolated from Europeans belong to hpEurope, including Basques in Spain, Russians and Kazakhs, and most isolates from east Asia belong to hpEastAsia (Supplementary Table 1). The results also confirmed that hpAfrica2, previously represented by only few isolates, is common in South Africa. Two new populations were identified and designated hpAsia2 and hpNEAfrica. hpAsia2 was isolated in northern India, Thailand, Bangladesh, the Philippines and elsewhere in southeastern Asia (Supplementary Fig. 1b). hpNEAfrica was predominant among isolates from Ethiopia, Somalia, Sudan and Nilo-Saharan speakers in northern Nigeria.

Matrices of pairwise  $F_{ST}$  (a measure of genetic differentiation between populations) between paired groups of samples from analogous geographic locations were strongly correlated (Mantel regression coefficient = 0.86,  $P < 0.001$ ) between *H. pylori* sequences and human microsatellite data<sup>16</sup>; 73% of human variation can be explained by a linear relationship with microbial variation (Supplementary Fig. 2). Thus, the geographic component of genetic diversity seems to be quantitatively comparable between *H. pylori* and humans, except that  $F_{ST}$  is considerably higher in *H. pylori*.

We next address evidence for a continuum of genetic ancestry in *H. pylori*. Whereas the assignments of individual isolates to populations are quite unambiguous with the no-admixture model in STRUCTURE<sup>17</sup>, its linkage model shows that the proportions of ancestry from the five

<sup>1</sup>Department of Molecular Biology, Max-Planck Institut für Infektionsbiologie, D-10117 Berlin, Germany. <sup>2</sup>Theoretical and Molecular Population Genetics Group, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. <sup>3</sup>Evolutionary Ecology Group, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. <sup>4</sup>Department of Statistics, University of Oxford, Oxford OX1 3SY, UK. <sup>5</sup>Génétique et Evolution des Maladies Infectieuses, UMR IRD-CNRS 2724, centre IRD de Montpellier, 911 Avenue Agropolis, BP 64501, 34394 Montpellier Cedex 05, France. <sup>6</sup>Department of Internal Medicine and Gastroenterology, University of Pretoria, Pretoria 0002, South Africa. <sup>7</sup>Department of Medicine—Gastroenterology, Baylor College of Medicine and Michael E. DeBakey VA Medical Center, Houston, TX 77030, USA. <sup>8</sup>Department of Microbiology, Donostia Hospital, 20014 San Sebastian, Spain. <sup>9</sup>Department of Laboratory Medicine, Lund University, SE22632 Lund, Sweden. <sup>10</sup>Medizinische Hochschule Hannover, Institut für Medizinische Mikrobiologie und Krankenhaushygiene, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany.

\*These authors contributed equally to this work.



**Figure 1 | Five ancestral populations in *H. pylori*.** **a**, DISTRUCT<sup>24</sup> plot of the proportions of ancestral nucleotides in 769 *H. pylori* isolates as determined by Structure V2.0 (linkage model)<sup>17</sup>. A thin line for each isolate indicates the estimated amount of ancestry from each ancestral population as five coloured segments (see **b** for colour code). The lines are grouped by (sub)population and ordered by ancestry. **b**, Neighbour-joining tree (black lines) of the relationships between and diversity within five ancestral populations calculated as described<sup>2</sup>. Circle diameters are proportional to

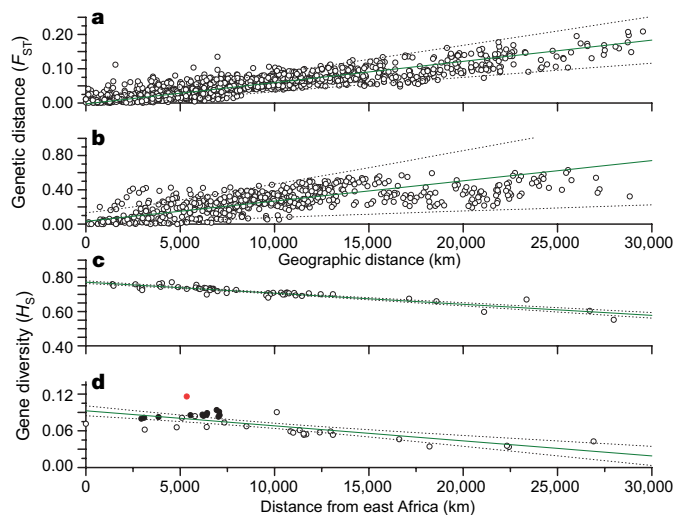
within-population genetic diversity and angles of filled arcs are proportional to the amount of ancestry attributable to each population among modern strains. Scale bar, 0.01. **c–g**, Spatial distribution of five geographic sources of ancestral nucleotides. Dark–light gradients show clinal declines in proportions of ancestral nucleotides by distance from a geographic centre. Colour-coding as in **b**. AE1, AE2, ancestral Europe 1 and 2, respectively. hspAmerind, hspMaori, hspEAsia are subpopulations of hpEastAsia; hspSAfrica and hspWAfrica are subpopulations of hpAfrica1.

ancestral sources almost form a continuum (Fig. 1a) between the five populations other than hpAfrica2, which is highly distinct. Similarly, although most concatenated sequences cluster according to their population assignment in a phylogenetic tree (Supplementary Fig. 3), their relatedness is also almost continuous, again with the exception of hpAfrica2. A nearly continuous distribution of the proportions of ancestry suggests localized admixture due to recombination. Such admixture would blur differences between initially distinct populations in close geographical proximity, and could potentially lead to strong signals of IBD in *H. pylori*, as observed in humans<sup>9–12</sup>.

Diversity was only poorly correlated with geographical sources in initial analyses, which might reflect noise due to recent human migrations plus horizontal transmission of *H. pylori* between ethnically distinct groups. Therefore, we excluded 147 isolates from obvious recent human migrants and their admixed descendants as well as 31 isolates whose population assignments were highly incongruent with their sources of isolation (horizontal transmission) (see Methods). The resulting ‘non-migrant’ data set (Supplementary Table 2) contained 532 *H. pylori* isolates and 1,405 polymorphisms. The results were compared with human diversity based on 783 autosomal microsatellites<sup>16</sup>. Similar to previous analyses<sup>11,12,14</sup>, 77% of the variance in  $F_{ST}$  between autosomal human markers from distinct geographic sources was accounted for by the shortest geographic distance along landmasses (Fig. 2a). For *H. pylori*, only 47% of the variance was accounted for by geographic distance (Fig. 2b,  $P \leq 0.001$ ), but this estimate rose to 72% when a standard conversion of genetic diversity was plotted against the logarithm of the geographic distance for the 442 haplotypes from geographic locations with at least 10 isolates (Supplementary Fig. 4). Thus, comparable proportions of the genetic diversity are due to IBD in *H. pylori* as in humans.

Genetic diversity within modern humans decreases with distance from east Africa, reflecting their recent African origin<sup>12–14</sup>; 85% of this decrease in diversity could be accounted for by distance from east

Africa (Fig. 2c). The non-migrant *H. pylori* data set also showed a similar trend (Fig. 2d,  $P \leq 0.001$ ) and 59% of the decrease in diversity could be accounted for by distance from east Africa. Unlike IBD, where trends with *H. pylori* might mimic those of humans without a joint demography, parallel decreases in diversity with distance from east Africa indicate close associations between the two. Simulations



**Figure 2 | Parallel geographic patterns of genetic diversity in humans and *H. pylori*.** **a, b**, Genetic distance in humans (**a**) and *H. pylori* (**b**) between pairs of geographic populations ( $F_{ST}$ ) versus geographic distance between the two populations. **c, d**, Average gene diversity in humans (**c**) and *H. pylori* (**d**) within geographic populations ( $H_S$ ) versus geographic distance from east Africa.  $R^2$  is 0.77 for **a**; 0.47 for **b**; 0.85 for **c**; and 0.59 for **d**. In **d**, samples that are predominantly composed of hpEurope isolates are indicated by filled circles, whereas the red circle identifies the sample from South Africa. Confidence intervals are indicated by dotted lines.

**Table 1 | Simulations for three hypothetical scenarios for the origin of *H. pylori*.**

Scenario	Estimated parameters					Model	
	Ancestral population ( $K_0$ )	Carrying capacity ( $K$ )	Migration ( $K^*m$ )	Growth rate ( $r$ )	Time (years)	$R^2$	AIC
East Africa	561 $\pm$ 182	203 $\pm$ 87	65 $\pm$ 6	0.73 $\pm$ 0.14	57,955 $\pm$ 3,748	0.68	83.3
South Africa	541 $\pm$ 214	185 $\pm$ 76	86 $\pm$ 8	0.68 $\pm$ 0.13	61,746 $\pm$ 4,436	0.57	88.1
East Asia	747 $\pm$ 207	467 $\pm$ 191	370 $\pm$ 169	0.70 $\pm$ 0.16	36,500 $\pm$ 6,728	0.02	101.5

Additional details are in Supplementary Information. The five estimated parameters are the size of the founder population ( $K_0$ ) and the carrying capacity of any subsequently colonized demes ( $K$ ), both of which are expressed as the number of effective strains that succeed in infecting the subsequent host generation.  $K^*m$  represents the number of effective migrants sent by any deme,  $r$  is the growth rate within demes in a logistic growth model, and time refers to the duration of the entire colonization process. The time estimates presented here include an additional 2,500 yr migration phase after the last deme had been colonized. Mean and standard deviations were computed for all simulations that fell within the 95% confidence interval. The fit of the models to the data are given by  $R^2$ , the amount of variance explained, and AIC, the Akaike information criterion.

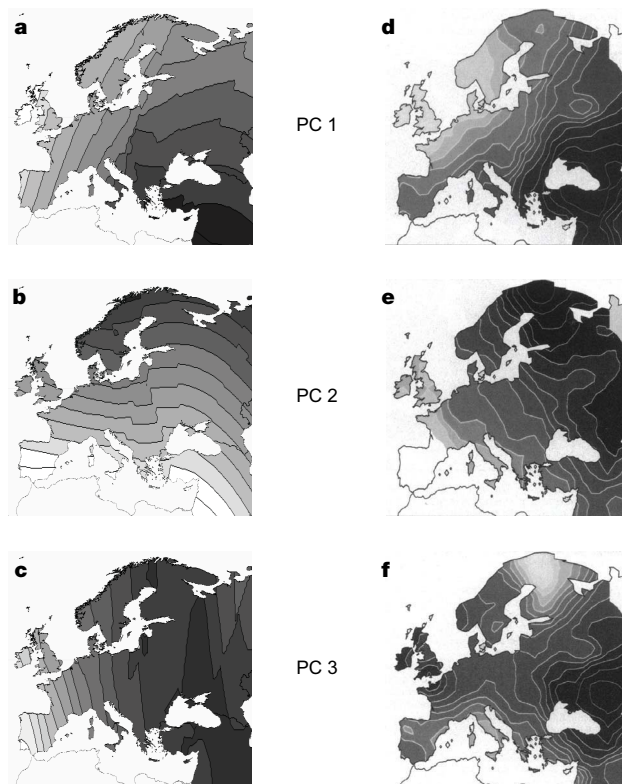
with human data indicated that they migrated from east Africa 56,000  $\pm$  5,500 yr ago<sup>14</sup>. Similar demographically explicit genetic simulations now indicate that an east African source for *H. pylori* is more probable than South Africa or China, and that *H. pylori* migrated from east Africa 58,000  $\pm$  3,500 yr ago (Table 1, Supplementary Table 3 and Supplementary Fig. 5; see Supplementary Information for details). Thus, we conclude that *H. pylori* accompanied anatomically modern humans during their migrations from Africa that have been estimated at 50,000–70,000 yr ago<sup>14,18–21</sup>. This implies that *H. pylori* was present in Africa before these migrations, suggesting that Africa is the source of both *H. pylori* and humans.

Are clines in genetic diversity truly contradictory to discrete clusters? Discrete clusters can be defined even within a perfectly continuous pattern owing to sampling artefacts<sup>10</sup>. But for *H. pylori*, geographic isolation is only a marginally better predictor of genetic differentiation than discrete clusters based on genetic similarity. Within a generalized linear model framework, assignment of populations to the six clusters defined by STRUCTURE explains 70% of the variance of pairwise  $F_{ST}$  for populations with 10 or more isolates, versus up to 72% by geographical distance. This situation resembles previous results for the geographic apportionment of human genetic diversity<sup>11,16</sup>, except that even when geography is first accounted for, clusters as defined by STRUCTURE still explain 11% of additional variance in genetic differentiation between *H. pylori* populations as compared with only 2% in humans. We therefore examined the modern geographic sources of the nucleotides associated with the five ancestral populations according to STRUCTURE (Fig. 1c–g). The spatial distribution of ancestral nucleotides indicated that ancestral Europe2 (AE2) originated in east Africa, AE1 in central Asia, ancestral EastAsia in east Asia and ancestral Africa1 and Africa2 in Africa. These data probably reflect extensive population expansions, subsequent to the global spread that accompanied migrations out of Africa, and may well reflect important episodes in human history during the Neolithic period and later.

If *H. pylori* were also a marker for human migrations at a more local scale, one would expect to find similar patterns between human and bacterial diversity within Europe, which was not one of the sources of ancestral nucleotides in *H. pylori*. Indeed, the first two principal components of spatial autocorrelation with hpEurope isolates (Fig. 3a, b) were very similar to those that had been obtained with human allozymes in classical work by Cavalli-Sforza and colleagues<sup>7</sup> (Fig. 3d, e) and the third principal component (PC) showed similar east–west clines. Such clines were originally interpreted as genetic signatures or indications of episodic migratory events<sup>7</sup>, although this interpretation has been questioned<sup>22</sup>. We note that the first principal component of the *H. pylori* data, PC1, is a cline from the southeast that correlates ( $R^2 = 0.35$ ,  $P < 0.01$ ) with the proportion of ancestry from AE2 (Supplementary Fig. 6a), which originated in northeastern Africa. For PC2, a cline from the northeast, the correlation ( $R^2 = 0.6$ ,  $P < 0.01$ ) is with AE1 (Supplementary Fig. 6b), which originated in Central Asia. These correlations show that in *H. pylori*, as previously suggested<sup>2</sup>, much of the spatial pattern observed in Europe can be attributed to admixture from different sources. It also supports the controversial hypothesis<sup>7</sup> that similar clines in humans are also due to waves of migration of genetically

distinct populations into Europe (demic diffusion), except that the spatial sources of ancestral nucleotides are assigned to northeastern Africa and Central Asia. We further conclude that there are highly striking, quantitative parallels in clines and IBD both at the global and the local scale between humans and *H. pylori*. These presumably reflect the dissemination of *H. pylori* by a variety of prehistoric human migrations, followed by admixture after horizontal transmission between human populations.

In this paper we have shown that the key patterns in the distribution of *H. pylori* genetic diversity mirror those of its human host. At a worldwide scale, we recovered similar patterns of isolation by distance, though absolute genetic differentiation is higher in *H. pylori*. As in humans, we observed a continuous loss of genetic diversity with increasing geographic distance from east Africa, the likely cradle of anatomically modern humans. Even at the more restricted scale of Europe, we largely recreated the classical clines described by Cavalli-Sforza and colleagues. Finally, simulations predict that *H. pylori* has spread from east Africa over the same time scale as anatomically modern humans. These extraordinary parallel population genetic patterns between *H. pylori* and its human host all demonstrate an old association predating the ‘out of Africa’ event<sup>4</sup>. The results further point to a scenario where *H. pylori* and human



**Figure 3 | Similar clinal gradients between principal components 1–3 from European *H. pylori* and humans. a–c, *H. pylori* concatenated sequences. d–f, Human allozymes. Panels d–f, reprinted with permission from citation 7 (copyright 1995 National Academy of Sciences, USA).**



populations have been evolving intimately ever since, with limited long-range transmission by horizontal infections.

## METHODS

**Bacterial isolates and sequencing.** The expanded *H. pylori* data set consists of 3,406 base pairs (bp) of unique, concatenated sequences of fragments of *atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI* and *yphC* from 769 *H. pylori* isolates (Supplementary Table 1). The data set includes 347 novel isolates in addition to data from 422 other strains that have been described previously<sup>2,3,23</sup>. The new bacteria were isolated from 25 additional ethnic sources in Asia (8 countries), Europe (4 countries, including Basques), Africa and the Middle East (9 countries) and South America (2 countries), for a total of 51 ethnic sources (Supplementary Table 1). The forward and reverse strands were sequenced as described<sup>1</sup>. Almost half (1,522 sites, 45%) of the nucleotides are polymorphic, resulting in a nucleotide diversity ( $\pi$ ) of 4.2% for the entire data set.

The non-migrant data set excluded bacteria that were isolated from the following migrant human populations: Europeans and Cape Coloureds from Cape Town; Mestizos from Colombia and Venezuela; Whites and African Americans from the USA; isolates in Thailand from Chinese or without ethnic association. hpAfrica2 isolates from Xhosa near Pretoria were excluded because they were a selective subset rather than a population-wide sample. The Philippines' samples were also removed because almost all bacterial populations were found there, probably owing to their colonial history. For isolates from Native Americans, only hspAmerind strains were considered non-migrant. The data set was further restricted to geographic samples with at least four isolates, to avoid statistical noise, which resulted in the elimination of all Jewish and Russian isolates and singletons from locations in China and Japan.

Other methods and details are to be found in Supplementary Information.

Received 4 October; accepted 22 December 2006.

Published online 7 February 2007.

1. Achtman, M. *et al.* Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.* **32**, 459–470 (1999).
2. Falush, D. *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585 (2003).
3. Wirth, T. *et al.* Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proc. Natl Acad. Sci. USA* **101**, 4746–4751 (2004).
4. Eppinger, M. *et al.* Who ate whom? Adaptive *Helicobacter* genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet.* **2**, e120 (2006).
5. Kersulyte, D. *et al.* Differences in genotypes of *Helicobacter pylori* from different human populations. *J. Bacteriol.* **182**, 3210–3218 (2000).
6. Dailidiene, D. *et al.* *Helicobacter acinonychis*: genetic and rodent infection studies of a *Helicobacter pylori*-like gastric pathogen of cheetahs and other big cats. *J. Bacteriol.* **186**, 356–365 (2004).
7. Piazza, A. *et al.* Genetics and the origin of European languages. *Proc. Natl Acad. Sci. USA* **92**, 5836–5840 (1995).
8. Suerbaum, S. & Michetti, P. *Helicobacter pylori* infection. *N. Engl. J. Med.* **347**, 1175–1186 (2002).

9. Relethford, J. H. Global patterns of isolation by distance based on genetic and morphological data. *Hum. Biol.* **76**, 499–513 (2004).
10. Serre, D. & Paabo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
11. Manica, A., Prugnolle, F. & Balloux, F. Geography is a better determinant of human genetic differentiation than ethnicity. *Hum. Genet.* **118**, 366–371 (2005).
12. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15942–15947 (2005).
13. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15**, R159–R160 (2005).
14. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237 (2006).
15. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, New Jersey, 1994).
16. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, e70 (2005).
17. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
18. Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
19. Ingman, M. & Gyllenstein, U. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. *J. Hered.* **92**, 454–461 (2001).
20. Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186 (2003).
21. Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl Acad. Sci. USA* **103**, 9381–9386 (2006).
22. Currat, M. & Excoffier, L. The effect of the Neolithic expansion on European molecular diversity. *Proc. Biol. Sci.* **272**, 679–688 (2005).
23. Momynaliev, K. T. *et al.* Population identification of *Helicobacter pylori* isolates from Russia. *Genetika* **41**, 1182–1185 (2005).
24. Rosenberg, N. A. DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We acknowledge the receipt of bacterial strains from A. van der Ende, M. J. Blaser, N. J. Saunders, R. J. Owen, F. Mégraud and sequences from K. T. Momynaliev and C. Kraft. We thank J. Goudet for providing a modified version of FSTAT able to deal with the large data set and help with R by K. -P. Pleissner. Grant support was from the German Federal Ministry for Education and Research (BMBF) in the framework of the PathoGenoMik Network (M.A., S.S.), the Biotechnology and Biological Sciences Research Council (F.B.), the Swedish Research council (T.W.) and Lund University Hospital (T.W.).

**Author Information** EMBL accession numbers for DNA sequences, AM413111–418360. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A. (achtman@mpiib-berlin.mpg.de), F.B. (fb255@mole.bio.cam.ac.uk) or S.S. (Suerbaum.Sebastian@mh-hannover.de).