Genomic evolution and transmission of *Helicobacter pylori* in two South African families

Xavier Didelot[a,1], Sandra Nell[b], Ines Yang[b], Sabrina Woltemate[b], Schalk van der Merwe[c], Sebastian Suerbaum[b,d,1]

[a] Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK [b] Institute of Medical Microbiology and Hospital Epidemiology, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany [c] Department of Liver and Biliopancreatic diseases, University Hospitals KU Leuven, Herestraat 49, B - 3000 Leuven, Belgium [d] DZIF - German Center for Infection Research, Hannover-Braunschweig Site, Carl-Neuberg-Str. 1, 30625 Hannover, Germany

Submitted to Proceedings of the National Academy of Sciences of the United States of America

***Helicobacter pylori* infects the stomachs of one in two humans and can cause sequelae that include ulcers and cancer. Here we sequenced the genomes of 97 *H. pylori* isolates from 52 members of two families living in rural conditions in South Africa. From each of 45 individuals, two *H. pylori* strains were isolated from the antrum and corpus parts of the stomach, and comparisons of their genomes enabled to study within-host evolution. In 5 of these 45 hosts, the two genomes were too distantly related to be derived from each other and therefore represented evidence of multiple infections. From the remaining 40 genome pairs, we estimated that the synonymous mutation rate was 1.38x10-5 per site per year, with a low effective population size within host probably reflecting population bottlenecks and immune selection. Some individuals showed very little evidence for recombination, whereas in others, recombination introduced up to a hundred times more substitutions than mutation. These differences may reflect unequal opportunities for recombination depending on the presence or absence of multiple infections. Comparing the genomes carried by distinct individuals enabled to establish probable transmission links. Transmission events were found significantly more frequently between close relatives, and between individuals living in the same house. We found however that a majority of individuals (27/52) were not linked by transmission to other individuals. Our results suggest that transmission does not always occur within families, and that co-infection with multiple strains is frequent and evolutionarily important despite a fast turnover of the infecting strains within-host.**

genome sequence | Helicobacter pylori | mutation | recombination | transmission

Introduction

Helicobacter pylori is a bacterial pathogen that infects the stomach of about half of the human worldwide population. It is often carried asymptomatically for decades, but can also cause severe complications (1). It was first discovered by Warren and Marshall (2) to be a causative agent of stomach inflammation and both gastric and duodenal ulcers, and has since also been recognized as the cause of approximately one in twenty of all human cancers (3). A very fruitful approach to study *H. pylori* has been the creation of a Multilocus Sequence Analysis (MLSA (4)) scheme specific to this species (5). MLSA revealed much about the evolutionary history and global population structure of *H. pylori*, including that it infected anatomically modern humans at least 100,000 years ago, that it accompanied its host out of Africa 60,000 years ago and that its current worldwide genetic variation reflects the human migrations that have happened since (6–8). This intimate relationship with humans combined with a fast rate of evolution make *H. pylori* a useful marker to trace the movements of human populations (9, 10).

A key evolutionary property of *H. pylori* is its very high recombination rate. This mechanism has been well studied through laboratory experiments (11–13). *In vivo*, extensive recombination was first detected by comparison of the phylogenetic signals of three gene fragments (14). MLSA studies also revealed high rates of admixture (6–10), although the relationships between strains sampled from different individuals are typically too complex to allow a complete reconstruction of the evolutionary events separating them. A simpler approach has been to compare isolates taken sequentially from the same patient to study within-host evolution. Such comparisons were first based on a handful of genes (15), later on extended panels of genes (16), and culminated with the use of whole genomes (17). These studies confirmed the prominent role played by recombination in the genomic evolution of *H. pylori*, with up to 40% of genes affected over three years of within-host evolution (17).

In spite of its high medical importance, many questions remain unanswered about both the evolution and epidemiology of *H. pylori*. In terms of genomic evolution, there is a need to precisely quantify mutation and recombination rates, to investigate the effect of immune selection, and to describe the frequency and evolutionary role played by co-infections with multiple strains. In terms of epidemiology, *H. pylori* is thought to be transmitted either by oral-oral or fecal-oral route between close relatives within families (18), but this hypothesis needs to be formally tested and it is unclear whether infection can occasionally come from other sources. In the population genetic studies described above (6–10), precise inference about fine-scale evolution and epidemiology is typically impossible because the individual strains are too complexly related, partly as a result of the high recombination rate. On the other hand, in the within-host evolution studies described above (15–17), genomic evolution becomes much clearer but the ability to study other processes such as transmission is completely lost.

Here we take an intermediate strategy between these two types of previous studies. We present a whole genome analysis of 97 *H. pylori* isolates from 52 members of two well-characterized extended families living in the same rural community in Ogies, Mpumalanga, South Africa. These individuals lived under comparable social circumstances with 100% of the households having a reticulated water supply, 87% flushing toilets and 98% of individuals having their own toothbrush. They have been previously studied with regards to the role of dental plaque as a reservoir of *H. pylori* and for co-infections (19, 20). *H. pylori* isolates from these individuals have been analysed based on the sequences of

Reserved for Publication Footnotes

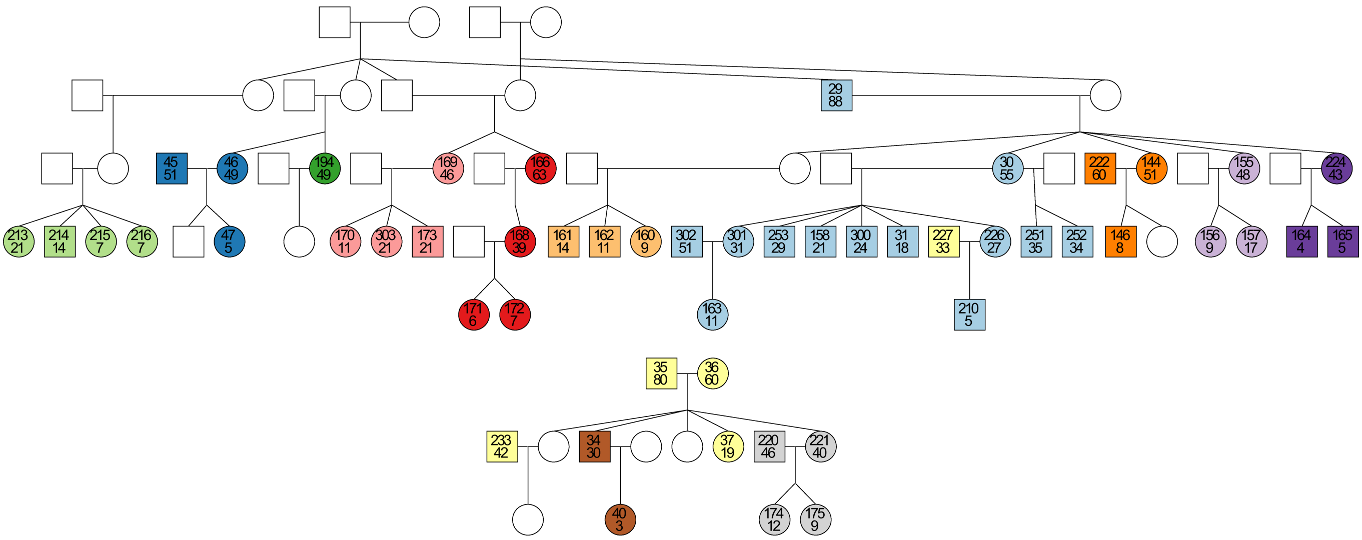

Fig. 1. Pedigrees of families 12 (top) and 13 (bottom). Males and females are represented by squares and circles, respectively. Each participant is labelled with a unique identifier number (top) and his or her age (bottom). Individuals not included in the study (because they were dead, unwilling to participate or did not carry *H. pylori*) are shown in white. Two participants are shown in the same colour if and only if they share the same house. There are a total of 13 unique colours representing the 13 different houses in which the participants live.

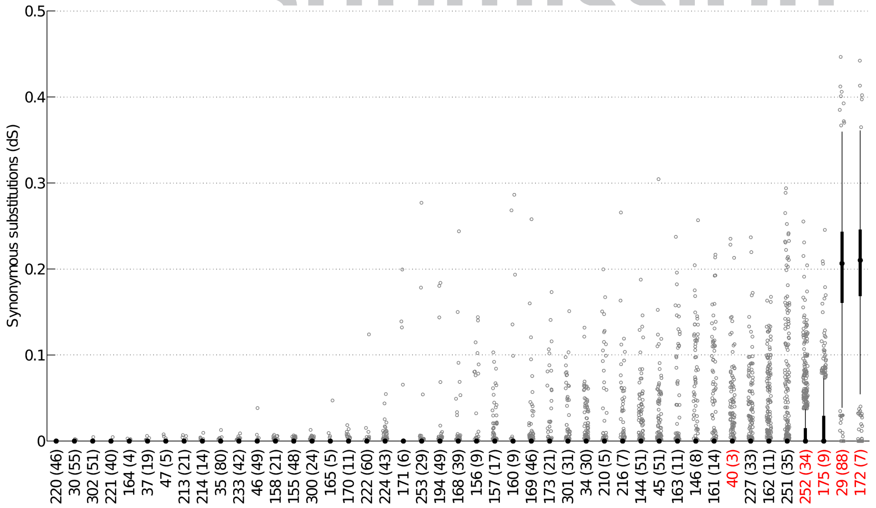

Fig. 2. Boxplots of the within-host synonymous distance (dS) for the 786 core genes of each of the 45 individuals in which two genomes were available. The individuals are sorted from left to right in increasing order of average dS across genes. Numbers in parentheses represent the age of the individual. The median dS is shown by a black dot, and the interquartile range (IQR) by a black rectangle. Boxplot whiskers have the standard maximum length of 1.5 times the IQR. In all but the last four individuals, the IQR spans from 0 to 0 and the black rectangle and whiskers are therefore not visible. Any gene with dS above or below the whiskers is shown as a grey open circle. The five individuals in red are the ones which were incompatible with evolution from a single infection.

three gene fragments (21) and using MLSA (22). These studies suggested a more complex mode of transmission than previously thought, although definite inference about who infected whom could not be reached on the basis of such small fractions of the genome. Here we sequenced the whole genomes of two isolates from most participants (one from the antrum and one from the corpus part of the stomach), so that within-host evolution can be studied through a comparison of the genomes carried by the same individual. Furthermore, transmission routes can also be investigated through comparisons of the genomes carried by different individuals.

Results

Novel genomic sequences. We used 454 technology (23) to fully sequence 97 genomes of *H. pylori* carried by 52 members of two South African families living in the same rural community. Family 12 included 42 participants from four generations and family 13 included 10 participants from three generations (Figure 1; Table S1). A total of 786 genes from reference strain 26695 (24) were found to be present in all 97 genomes, with a concatenated length of 709 kbp, which represents approximately 200 times more data per isolate than MLSA (Table S2, Figure S1). All the analyses presented below are based on these core genes. The distribution across core genes of the ratio of substitution rates at non-synonymous and synonymous sites dN/dS (Table S2) had a mean of 0.14 and a 95% central range of 0.02-0.34, indicating that genes were subject to varying levels of purifying selection. There were three outliers to this distribution of dN/dS across genes: HP0411 (dN/dS=2.50), HP1211 (dN/dS=1.01) and HP1145 (dN/dS=0.90). These three genes encoded unknown hypothetical proteins (24). These results are consistent with previously reported variations in the selective pressures acting between as well as within the genes of *H. pylori* (17, 25, 26). To guard against these variations, as well as against the relative strength of selection across time frames (16, 27) and the interaction between recombination and selection (28), we only used the synonymous substitutions in the remainder of the analysis.

Within-host evolution. There were 45 individuals for whom two *H. pylori* genomes (one from the antrum and one from the corpus part of the stomach) were available, and we measured the synonymous distance dS for each core gene of each pair of genomes (Figure 2). Based on these distances, we tested whether

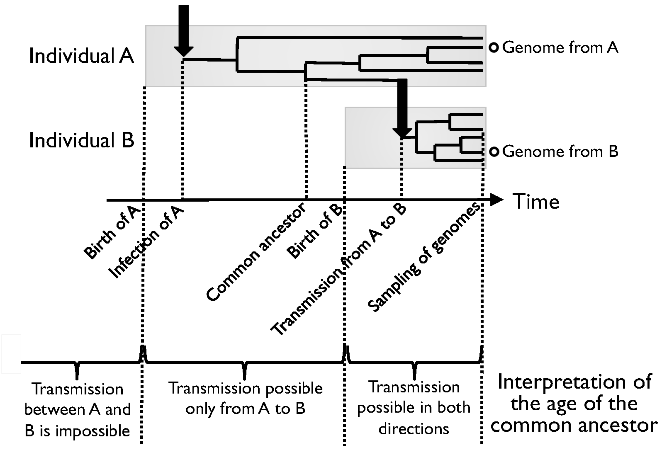
Individual A
Individual B
Genome from A
Genome from B
Time
Birth of A
Infection of A
Common ancestor
Birth of B
Transmission from A to B
Sampling of genomes
Transmission between A and B is impossible
Transmission possible only from A to B
Transmission possible in both directions
Interpretation of the age of the common ancestor

Fig. 3. Illustration of how the age of the common ancestor of a pair of genomes isolated from two distinct hosts is informative about the possibility of transmission between these two hosts. The top part shows an example of transmission from individual A to individual B. Within-host diversification is shown by the branching process, and transmission events are represented by the black arrows. Only one genome from each individual is considered at a time, and it is possible to estimate the date of their common ancestor. Because in this example the common ancestor is after the birth of A and before the birth of B, transmission from A to B is possible but not transmission from B to A. The bottom part of the figure gives the full scale of interpretation of possible dates for the common ancestor in term of transmission between individuals A and B.

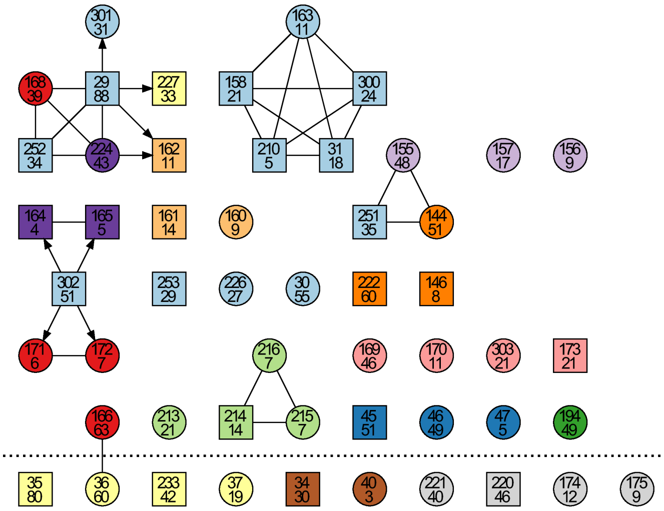

Fig. 4. Results of the transmission analysis. Individuals are represented using the same symbols, labels and colours as in Figure 1. The members of family 12 are shown above the dotted line whereas the members of family 13 are shown on the last row, below the dotted line. An arrow between two individuals indicates unidirectional possibility of transmission from a donor (tail of the arrow) to a recipient (head of the arrow). An undirected link between two individuals indicates that transmission is possible in both directions.

the two genomes were similar enough to have originated from the same infection. Pairs from 40 individuals were found to be compatible with a model of within-host evolution, where the differences between the genomes are explained by mutation and recombination events taking place during diversification from a common ancestor which postdates a single colonization event. Amongst the five individuals for which within-host evolution was impossible (highlighted in red in Figure 2), four had pairs of genomes with large numbers of non-identical genes therefore indicating multiple infection with at least two separate strains (individuals 29, 172, 175 and 252). In the last individual who was found to be incompatible with the within-host evolutionary model (individual 40), the two genomes had many identical genes indicating that they were related (Figure 2), but this individual was only 3 years old which did not leave enough evolutionary time to explain the high number of mutation and recombination events observed in other genes. This individual was therefore most probably infected twice with variants of the same strain, possibly from the same transmission donor, or have received multiple variants in the course of a single transmission.

Based on the pairs of genomes of the 40 individuals who were found to be compatible with within-host evolution, we estimated the synonymous mutation rate to be 1.38x10^{-5} per site per year, with a 95% credibility interval ranging from 9.14x10^{-6} to 1.85x10^{-5}. This is in good agreement with previous genomic estimates based on serial isolation of *H. pylori* from the same host (17), and one to two orders of magnitude higher than estimates in other bacterial species (29). The time to the most recent common ancestor (TMRCA) for each patient was typically of the order of only a few years, with an average of 3.61 years (Table S3). This was true even for older individuals, which was unexpected since our prior model stated that the TMRCA was equally likely to take any value between 0 and the age of the host (see Methods).

Recombination is a major driver of diversification in some individuals only. The ratio r/m (30, 31) of the rates at which recombination and mutation introduce substitutions was found to vary widely from one individual to another. Some individuals (e.g. 233, 158, 155, 300 in Table S3) had very low r/m values, potentially even equal to zero which would represent purely clonal evolution. In 14 individuals, no single gene was found to have a posterior probability of being recombined higher than 95% (Table S3). On the other hand, some individuals (e.g. 146, 171, 160, 161 in Table S3) had very high r/m values, similar or even higher than previously reported based on longitudinal isolates comparisons (16, 17). This result indicates that the effect of recombination varied significantly from one host to another.

Where recombination had taken place, the synonymous distance between donor and recipient was equal to 6.3% on average, with a central 95% range from 0.8% to 16.7% which covers the full span of pairwise distances between unrelated strains (6). This result reflects the high promiscuity of *H. pylori*, and contrasts for example with *Salmonella enterica*, *Escherichia coli* or *Bacillus cereus* where recombination was found to happen preferentially between members of the same lineage (32–34). Across all 40 pairwise comparisons of genomes corresponding to within-host evolution, the number of recombinant genes (i.e. with posterior probability of recombination above 95%) was found to vary from 0 to 95, for a total of 665 recombination events (Table S3). For 432 (65%) of these we were able to establish which of the two alleles was the recombinant one. We searched for putative origins of these recombinant alleles amongst all genomes, but only found possible donors for 54 (12.5%) of the imports, as summarized in Table S4. Only three instances were found with more than two genes putatively transferred from the same donor to the same recipient: these were from host 161 to 162 (10 genes), from 166 to 144 (7 genes) and from 161 to 216 (6 genes).

Transmission analysis. Having described the evolutionary dynamics of mutation and recombination from the relatively simple setting provided by within-host comparisons of genomes, we can now turn to the question of transmission between hosts. By measuring the time to the most recent common ancestor (TMRCA) of two genomes from two distinct individuals, it is possible to determine who may have infected whom, assuming that no genomic variation from the infector is transmitted to the infected (35), Transmission from host A to host B is only possible if this TMRCA is smaller than the age of host A (Figure 3). The results of this transmission analysis are shown in Figure 4. The majority of the individuals (27/52) were completely unlinked, meaning that they were neither donor nor recipient of any transmission event. In family 13, the only link concerned individual 36 who could have infected or been infected by individual 166 of family 12. This transmission link between the two families may have been established via individual 227 who is a member of family 12 but lives in the same house as 36 and other members of family 13 (Figure 1). Individual 227 may have acted as intermediate in the transmission between 166 and 36, even if he was not found to carry the same strain (he could either have lost it since or it could have been unsampled).

Several transmission clusters were detected in family 12, including one cluster made of three siblings (214, 215, 216) living in the same house, one cluster made of two siblings and their nephew (144, 155, 251) living in three separate houses, a cluster where one individual (302) infected two young sisters (171 and 172) and two young brothers (164 and 165), spanning three different houses, and one cluster made of three brothers (158, 300, 31) and their nephew (210) and niece (163), all of which live in the same house. Finally, the largest cluster was made of seven individuals (29, 301, 227, 162, 224, 252, 168) spanning four houses, with at its centre the oldest participant (29 who was 88 years old). Amongst all these transmission clusters, there was only one possible instance of transmission from parent to offspring (from 29 to 224).

The inferred patterns of transmission between individuals (Figure 4) were reconstructed solely on the basis of the homology of the *H. pylori* genomes they carry, and the age of the individuals. We can therefore compare these results with other epidemiological variables that were not used in the analysis, such as the degree of relatedness between individuals or the memberships of households (Figure 1). We found a strong positive association between transmission and kinship coefficient (r=0.22, p=5x10^{-6}, simple Mantel test), indicating that transmission happened more often between close relatives. We also found a similarly strong positive association between transmission and house sharing (r=0.24, p=8x10^{-6}; simple Mantel test). Since relatedness and house sharing are strongly correlated (Figure 1), we tested the association of each variable while controlling for the other in order to disentangle their effects. We found a positive association of transmission with kinship when controlling for house sharing (r=0.10, p=7.9x10^{-3}, partial Mantel test) and a positive association of transmission with house sharing when controlling for kinship (r=0.12, p=3.2x10^{-3}, partial Mantel test). These results indicate that both kinship and house sharing have independent effects of roughly the same strength on transmission patterns.

Discussion

Variation in recombination rates between hosts. *H. pylori* is often described as a highly recombinogenic species of bacteria (6, 14–17). Our study confirms that recombination can indeed be a potent force of genomic diversification, with several examples of within-host evolution where the ratio r/m of the effects of recombination and mutation may be over a hundred (Table S3). However, we also found that in some infected individuals, recombination has played a much more modest role if any. A first explanation for these differences could be that some lineages of *H. pylori* have higher rates of recombination than others. Such variations in recombination rates have been previously described between lineages of *Salmonella enterica* (32), *Escherichia coli* (36) or *Chlamydia trachomatis* (37). For *H. pylori*, laboratory experiments have shown that recombination rates depended significantly on the combination of recipient and donor strains (11, 13). Here however there seemed to be no relationship between infecting strain and effect of recombination. For example, the three individuals 155, 144 and 251 have infected each other (Figure 4) but the effect of recombination was found to vary greatly between them (with r/m mean estimates of 0.5, 13.8 and 36.6 respectively; Table S3). Another likely important source of variation in the effect of recombination is the presence or absence of multiple infections. Individuals with low r/m could correspond to individuals that were not multiply infected (so that recombination only happened between members of the same strain and therefore does not have much effect) whereas individuals with high r/m would be the ones where multiple infections were present. Our observation of strong variation of the rate of *in vivo* recombination, even within one community reconciles previous controversial reports about the frequency of recombination in *H. pylori*, which included both very high (15, 17) as well as very low (38) values.

Multiple *H. pylori* infections. Infections with multiple strains of *H. pylori* have been reported in several studies, but their true incidence is unknown because analysis of multiple isolates from one individual is rarely performed (39, 40). A recent study found that infections with multiple strains of *H. pylori* were very common in India (41). We found definite evidence for multiple infections in 5 out of 45 individuals, where the genome sequences of the *H. pylori* strains isolated from antrum and corpus shared only few identical genes. Table S3 shows the results for the remaining 40 individuals whose paired *H. pylori* genomes were highly related, such that for these individuals there was no direct evidence that multiple infections were present. It is however likely that some of them also hosted multiple infections that were not detected because only two isolates were studied in each case. Even if two infections were present in a host in equal proportions and complete mixture (giving the best chance to sample both), then the probability to take one isolate from each is only 50%. This would suggest that at least half of the multiple infections were undetected, giving an estimate for the total number of multiply infected individuals of at least 10 out of 45 (22%). It seems also likely that for some individuals, multiple infections were present in the past and acted as a source of recombined material, but had been removed by the time of sampling.

Diversity bottleneck and immune selection. In many individuals, the antrum and corpus genomes only differed at few genes (Figure 2). Consequently, the TMRCA of pairs of genomes from the same infection were relatively low, with an average of 3.61 years (Table S3). This result is in good agreement with the previous observation that the genetic distance between serial isolates from a same host is correlated with the time separating the two isolations but uncorrelated with the age of the individuals (16). One explanation for these low TMRCA would be that colonization with *H. pylori* has been recent for most individuals, but this would go against the generally accepted idea that acquisition happens in early childhood (1, 18) as well as the high proportion of individuals who were found to be carriers of *H. pylori* in this setting (21, 22). More probably, this observation could be the result of strong genetic drift (42), for instance population bottlenecks caused by the selective pressure of the human immune system (43). Assuming an average of one cell division per day (44), the average TMRCA of 3.61 years would imply an effective population size of *H. pylori* equal to N_e=1318. This quantity indicates how quickly polymorphism is lost in the population. For example, in a multiple infection by two strains present in equal proportions (giving them the best chance to both persist for a long time) the expected time until one or other strain was lost would be 5 years (45). The two oldest individuals where multiple infections were detected (individuals 252 and 29 who were 34 and 88 years old respectively) are therefore unlikely to represent acquisition of multiple strains in early childhood. Instead, they probably reflect that in highly endemic regions, even individuals who are already colonized are exposed to infection with new strains. Treatment with antibiotics could be another reason for the observed diversity bottlenecks. However, antibiotics can only obtained on prescription in South Africa, and use of antibiotics in the Ogies community is low. Only five of the individuals for whom antrum-corpus strain pairs were available had reported recent antibiotic use, and there was no correlation with TMRCA. We thus conclude that antibiotics are an unlikely major reason for the observed low average TMRCA.

Missing transmission links. In order to reconstruct putative transmission links between individuals, we compared the genomes they carried whilst accounting for the within-host evolution that could have happened before and after the transmission event (Figure 3). This approach revealed several instances of transmission, significantly over-represented between closely related members of the same family or between individuals living in the same house (Figure 4). We note that our genome-based approach identified far more inferred transmission links than a previous MLSA approach, which used the carriage of strains with an identical ST as evidence of likely recent transmission, and connected 14 individuals from the two families (22). Interestingly, all transmission links suggested by the “identical ST” criterion were supported by the whole genome approach, indicating that due to the rapid diversification of *H. pylori*, ST identity is in fact a highly specific criterion for a transmission link. However, our current approach linked 25 individuals, and provided support for a direction of transmission for seven links, which is not possible using MLSA alone.

However, there remained a majority of individuals (27/52) which could have acted neither as donor nor recipient of any transmission event. Similarly, for a large proportion (378/432) of the genes that were found to have been imported through recombination during within-host evolution, we were unable to identify a possible origin amongst the 97 genomes we sequenced. These two observations, that donors of both transmission and recombination events could not be identified, are in good agreement with each other. They both suggest that some of the strains circulating presently or in the past have not been sampled in this study.

A first explanation may be that transmission of *H. pylori* sometimes happens outside of the familial context, and therefore members of other families living in the same South African rural community would have to be studied to find the missing links. The fact that we found a putative transmission link between members of the two families (between hosts 166 and 36, cf Figure 4) confirms the validity of this hypothesis. However, since transmission was found to predominantly happen between relatives and inhabitants of the same houses, the influence of other families seems insufficient to explain on its own the large numbers of missing transmission and recombination links. A second explanation would be that multiple infections are common, and we have discussed above the evidence supporting this claim. This would imply that not all infections have been sampled in the participating individuals, partly because only two genomes were sampled from each of them but also because they may have hosted other strains in the past that were not present anymore by the time of the endoscopy.

These two hypotheses, transmission outside of the familial setting and frequent turnover of infecting strains, are not mutually exclusive and could together explain the large fraction of missing donors of transmission and recombination. They represent the two facets of what needs to be investigated to fully understand any infectious disease: within-host evolution on one hand, and transmission from host to host on the other. Whole-genome sequencing holds great promise to elucidate these processes, not just for *H. pylori* as we have demonstrated here, but also for many other microbial pathogens (29). To fully exploit this potential requires to consider within- and between-host dynamics jointly, since they can only be understood in light of each other.

Materials and Methods

Genome sequencing and assembly. Draft genome sequences of *H. pylori* isolates from families 12 and 13 from Ogies, Mpumalanga, South Africa (19–22) were obtained using Roche 454 FLX technology (23). Library preparation was either done according to GS FLX General Library Preparation Method Manuals for FLX Chemistry or FLX Titanium chemistry (overall 26 isolates of family 12). 54 isolates of family 12 and 19 isolates of family 13 were prepared according to the Rapid Library Preparation Method Manual. Emulsion PCR and 454 pyrosequencing were performed following the manufacturer's instructions. The genomes of two isolates (antrum isolate of individual 303 and corpus isolate of individual 174) did not pass quality control, leaving a total of 97 genomes: 79 genomes from 42 members of family 12 and 18 genomes from 10 members of family 13 (Table S1; Figure 1). The genomes were assembled *de novo* using the Roche GS De Novo Assembler (version 2.6), resulting in an average of 60 contigs per genomes (with a minimum of 26 and a maximum of 188; Table S1).

Identification of core genes and synonymous polymorphisms. The annotation of the previously sequenced reference genome *H. pylori* strain 26695 contains 1590 predicted coding sequences (24). For each of these genes and each of our 97 genomes, we used BLAST (46) to look for homologs, following a similar approach to the one implemented by BIGSdb (47). If the best BLAST hit of a gene against a genome covered at least 90% of the positions of the query sequence, the gene was considered to be found. This is a conservative approach to finding gene homologs since it does not allow long indels or gene splits at contig ends. We found that 786 genes (49%) were present amongst all 97 genomes (Table S2). These core genes ranged in length from 201bp up to 3636bp, with a mean of 902bp and a total concatenated length of 709,155bp, which represents 42.5% of the length of the 26695 genome (24). For each core gene, a query-anchored alignment was produced, so that each column of the alignment corresponded exactly to a position of the gene sequence in the reference genome. Consequently, our data was robust to insertions or deletions relative to the reference which may be due to the known tendency of 454 to produce homopolymer frameshift errors. The core genes are well distributed around the reference genome 26695 except for a few regions of very low GC content (Figure S1). All the analyses presented are based on these core genes so that comparisons between genomes are always based on exactly the same data. Synonymous and non-synonymous polymorphisms were distinguished using the method of Nei and Gojobori (48). The ratios dN/dS of rates of non-synonymous and synonymous mutations were computed for each gene (Table S2). Only synonymous substitutions were used in the analyses of within-host genomic evolution and host-to-host transmission.

Evolutionary model. Let us consider a pair of genomes. Let t denote the time to their most recent common ancestor, so that in evolutionary terms $2t$ separate the two genomes. During this time, each gene indexed j may have recombined (with probability 1-exp(-ρt) where $\rho/2$ is the recombination rate) or evolved clonally (with probability exp(-ρt)). If the gene j evolved clonally, it has accumulated a number of mutations distributed as Binomial(L_j, $t\theta_s$) where L_j is the number of synonymous sites in gene j, and $\theta_s/2$ is the synonymous mutation rate. If the gene j recombined, it would have acquired a number of substitutions distributed as Binomial(L_j, ν) where ν is the distance between donor and recipient of the recombination event, which is distributed as Beta(α, β). This evolutionary model is related to the ClonalFrame model (49), but is more general in that ν takes a distribution instead of being a constant, as previously proposed (16). The parameters ρ and t are assumed to be specific to each pairwise comparison, which represents another generalization of the model allowing to capture variations in the recombination rate. The parameters θ_s, α and β are identical across comparisons. Unless otherwise stated, the prior distribution on any parameter is improper uniform over $(0,\infty)$.

Analysis of within-host evolution. There were 45 individuals for whom two genomes were sequenced (one isolated from the antrum and one from the corpus), and the synonymous distance dS was measured for each gene of these paired genomes (Figure 2). For a given individual indexed i who is of age a_i, if we assume that the two genomes are descended from the same ancestor within the host (i.e. are part of the same infection), then let t_i denote the time to this common ancestor, which takes value between 0 (in the extreme case where the two genomes separated from each other just before isolation) and a_i (in the opposite extreme case where the individual was infected just after birth and the common ancestor dates from then). To reflect our initial ignorance about t_i, we assume a uniform prior between 0 and a_i for this parameter. The two genomes are assumed to be related according the model described in the previous paragraph, and inference was performed for all parameters using a Monte-Carlo Markov Chain (MCMC), initially considering that all pairs of genomes from the same individual had evolved within-host. Posterior-predictive distributions (50) on the number of substitutions between pairs of genomes were used to assess the fit of this model for each individual. In five individuals (29, 40, 172, 175 and 252), the number of differences was found to be incompatible with within-host evolution, which means that the two genomes result from two separate infection events. The MCMC was rerun with these individuals excluded in order to estimate all parameters (Table S3 and Table S5).

Analysis of recombination. In the evolutionary model described above, the relative effect of recombination and mutation r/m (30, 31) is equal to $\rho\alpha/(\theta(\alpha+\beta))$ and this value is reported for each host in Table S3. For all recombination events with posterior probability above 95%, we attempted to establish which of the two alleles was the recombinant one by comparing them with the allele carried by the closest overall relative of the two genomes. If the distance of one allele to that of the closest relative was at least twice as high as that of the other allele, then we considered that the former was the recombinant allele. For each imported gene we searched for an exact match in another genome, which would therefore represent a possible source of recombination (32, 51).

Analysis of transmission. For each pair of genomes from different patients, we estimated the time t of their most recent common ancestor using the model above and setting the value of the global parameters θ_s, α and β to their posterior mean from the within-host analysis above. Transmission from individual A to individual B was assessed to be possible only if the lower bound of the 95% credibility interval of t was smaller than the age a_A of individual A. In other words, transmission from A to B is possible only if the age of the common ancestor of the genomes carried by A and B is smaller than the age of A, so that this common ancestor could have existed in individual A (Figure 3). This is assuming that only a single genomic variant is transmitted from A to B, or in other words that there is a strong bottleneck of diversity at the point of transmission. The correlation of transmission links (Figure 4) with kinship coefficients and household memberships were measured using Mantel tests. Partial Mantel tests with permutation of residuals were used to test the association of transmission with kinship while controlling for housesharing and vice-versa. These Mantel tests were performed only in family 12 since no transmission events were found within family 13, and they were conducted using a million random permutations in the statistical software zt (52). The analyses of within-host evolution, recombination and transmission described above were performed using Matlab code that was developed specifically for the purpose of this study, and which is available from the authors upon request.

Acknowledgements.

The authors thank Mark Achtman, Daniel Falush and Christine Josenhans for helpful comments on an earlier version of this manuscript, and Friederike Kops and Birgit Brenneke for excellent technical assistance. This study was supported by grants SFB 900/A1 an SFB 900/Z1 from the German Research Foundation (DFG) to S.S.

1. Suerbaum S, Josenhans C (2007) *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat Rev Microbiol* 5:441–52.
2. Marshall BJ, Warren JR (1983) Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* 321:1273–1275.
3. Parkin DM (2006) The global health burden of infection-associated cancers in the year 2002. *Int J Cancer* 118:3030–44.
4. Maiden MC et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 95:3140–5.
5. Achtman M et al. (1999) Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol Microbiol* 32:459–70.
6. Falush D et al. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* 299:1582–5.
7. Linz B et al. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445:915–918.
8. Moodley Y et al. (2012) Age of the Association between *Helicobacter pylori* and Man. *PLoS Pathog* 8:e1002693.
9. Wirth T et al. (2004) Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: Lessons from Ladakh. *Proc Natl Acad Sci USA* 101:4746–4751.
10. Moodley Y et al. (2009) The peopling of the Pacific from a bacterial perspective. *Science* 323:527–530.
11. Kulick S et al. (2008) Mosaic DNA imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*. *PLoS One* 3:e3797.
12. Lin EA et al. (2009) Natural transformation of *Helicobacter pylori* involves the integration of short DNA fragments interrupted by gaps of variable size. *PLoS Pathog* 5:e1000337.
13. Moccia C et al. (2012) The nucleotide excision repair (NER) system of *Helicobacter pylori*: Role in mutation prevention and chromosomal import patterns after natural transformation. *BMC Microbiol* 12:67.
14. Suerbaum S et al. (1998) Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA* 95:12619–24.
15. Falush D 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.
16. Morelli G et al. (2010) Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet* 6:e1001036.
17. Kennemann L et al. (2011) *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci USA* 108:5033–5038.
18. Goh K-L, Chan W-K, Shiota S, Yamaoka Y (2011) Epidemiology of *Helicobacter pylori* infection and public health implications. *Helicobacter* 16 Suppl 1:1–9.
19. Olivier BJ et al. (2006) Absence of *Helicobacter pylori* within the Oral Cavities of Members of a Healthy South African Community. *J Clin Microbiol* 44:635–636.
20. Fritz EL, Slavik T, Delport W, Merwe SW Van Der (2006) Incidence of *Helicobacter felis* and the Effect of Coinfection with *Helicobacter pylori* on the Gastric Mucosa in the African Population. *J Clin Microbiol* 44:1692–1696.
21. Delport W, Cunningham M, Olivier B, Preisig O, Van der Merwe SW (2006) A population genetics pedigree perspective on the transmission of *Helicobacter pylori*. *Genetics* 174:2107–18.
22. Schwarz S et al. (2008) Horizontal versus familial transmission of *Helicobacter pylori*. *PLoS Pathog* 4:e1000180.
23. Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nat Biotechnol* 26:1117–24.
24. Tomb JF et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388:539–47.
25. Duncan SS et al. (2013) Comparative Genomic Analysis of East Asian and Non-Asian *Helicobacter pylori* Strains Identifies Rapidly Evolving Genes. *PLoS One* 8:e55120.
26. Olbermann P et al. (2010) A Global Overview of the Genetic and Functional Diversity in the *Helicobacter pylori* cag Pathogenicity Island. *PLoS Genet* 6:17.
27. Rocha EPC et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239:226–35.
28. Castillo-Ramírez S et al. (2011) The Impact of Recombination on dN/dS within Recently Emerged Bacterial Clones. *PLoS Pathog* 7:e1002129.
29. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* 13:601–612.
30. Feil EJ et al. (2001) Recombination within natural populations of pathogenic bacteria : Short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* 98:182–187.
31. Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 3:199–208.
32. Didelot X et al. (2011) Recombination and Population Structure in *Salmonella enterica*. *PLoS Genet* 7:e1002191.
33. Didelot X, Lawson DJ, Darling AE, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* 186:1435–49.
34. Didelot X, Meric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13:256.
35. Didelot X et al. (2012) Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* 13:R118.
36. Wirth T et al. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 60:1136–51.
37. Joseph SJ et al. (2012) Population Genomics of *Chlamydia trachomatis*: Insights on Drift, Selection, Recombination and Population Structure. *Mol Biol Evol* 29:3933–46.
38. Lundin A et al. (2005) Slow Genetic Divergence of *Helicobacter pylori* Strains during Long-Term Colonization. *Infect Immun* 73:4818–4822.
39. Taylor NS et al. (1995) Long-term colonization with single and multiple strains of *Helicobacter pylori* assessed by DNA fingerprinting. *J Clin Microbiol* 33:918–23.
40. Kersulyte D, Chalkauskas H, Berg DE (1999) Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Mol Microbiol* 31:31–43.
41. Patra R et al. (2012) Multiple Infection and Microdiversity among *Helicobacter pylori* Isolates in a Single Host in India. *PLoS One* 7:e43370.
42. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205.
43. Young BC et al. (2012) Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci USA* 109:4550–4555.
44. Dumrese C et al. (2009) The secreted *Helicobacter* cysteine-rich protein A causes adherence of human monocytes and differentiation into a macrophage-like phenotype. *FEBS letters* 583:1637–1643.
45. Kimura M, Ohta T (1969) The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63:701–709.
46. Altschul SF et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
47. Jolley KAA, Maiden MCJC (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595.
48. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–26.
49. Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–66.
50. Gelman A, Meng X, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sinica* 6:733–807.
51. Didelot X, Barker M, Falush D, Priest FG (2009) Evolution of pathogenicity in the *Bacillus cereus* group. *Syst Appl Microbiol* 32:81–90.
52. Bonnet E, Van de Peer Y (2002) zt: a software tool for simple and partial Mantel tests. *J Stat Softw* 7:1–12.