Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age

Daniel Falush*, Christian Kraft[†], Nancy S. Taylor[‡], Pelayo Correa[§], James G. Fox[‡], Mark Achtman*, and Sebastian Suerbaum[†¶]

*Max-Planck Institut für Infektionsbiologie, Schumannstrasse 21/22, 10117 Berlin, Germany; [†]Institut für Hygiene und Mikrobiologie, Universität Würzburg, Josef-Schneider Strasse 2, 97080 Würzburg, Germany; [‡]Division of Comparative Medicine, Massachusetts Institute of Technology, Cambridge, MA 02139; and [§]Louisiana State University Medical Center, New Orleans, LA 70112

Edited by Henry C. Harpending, University of Utah, Salt Lake City, UT, and approved October 16, 2001 (received for review July 28, 2001)

The bacterium *Helicobacter pylori* colonizes the gastric mucosa of half of the human population, resulting in chronic gastritis, ulcers, and cancer. We sequenced ten gene fragments from pairs of strains isolated sequentially at a mean interval of 1.8 years from 26 individuals. Several isolates had acquired small mosaic segments from other *H. pylori* or point mutations. The maximal mutation rate, the import size, and the frequency of recombination were calculated by using a Bayesian model. The calculations indicate that the last common ancestor of *H. pylori* existed at least 2,500–11,000 years ago. Imported mosaics have a median size of 417 bp, much smaller than for other bacteria, and recombination occurs frequently (60 imports spanning 25,000 bp per genome per year). Thus, the panmictic population structure of *H. pylori* results from very frequent recombination during mixed colonization by unrelated strains.

Bayesian model | horizontal genetic exchange | genomic flux | evolution

Helicobacter pylori colonizes the stomachs of more than one half of the world population. It is transmitted within families and occasionally from other sources (1). The sequence diversity of its housekeeping genes exceeds that of most other bacteria (2) and is associated with an exceptionally high frequency of distinct alleles. *H. pylori* sequences have a uniquely high homoplasy ratio (3), an indirect measure of genetic shuffling. These observations are thought to result from horizontal genetic exchange during mixed colonization by unrelated strains. Geographic structure exists within *H. pylori* and sequences are less related between isolates from different continents than between isolates from single countries (4).

Multiple recombinants have been isolated from one individual (5), but the frequency and other basic parameters of recombination have not yet been estimated. The age of *H. pylori* is unknown, and standard methods for estimating age based on sequence diversity cannot be used without an estimated molecular clock rate.

In this report we present data on the frequency of imports and mutations within paired sequential isolates from patients from two geographical areas. A model was developed that estimates recombination size, recombination rate, and mutation frequency from such data. This model can be used to set lower limits on the age of *H. pylori* and other bacteria with frequent recombination.

Materials and Methods

Bacterial Isolates. Single colonies of *H. pylori* were isolated from sequential biopsies taken during two clinical trials. Louisiana (6): A clinical treatment trial in the 1980s in New Orleans involving patients at high risk of infection, most of whom were black. For those patients with multiple sequential isolates, the earliest and latest were used. Colombia (7): A trial addressing the effects of chemoprevention on the progression of precancerous lesions among Mestizos with multifocal atrophic gastritis in Narino in the Colombia Andes.

Nucleotide Sequencing and Sequence Analysis. Core fragments of seven housekeeping genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, *yphC*) and three virulence associated genes (*flaA*, *flaB*, *vacA*) (Fig. 1) were sequenced as described (3, 4). Regions flanking the core fragments were sequenced by using additional oligonucleotide primers (details available on request). All sequences have been deposited in the GenBank database (accession nos. AJ418065–AJ418366).

The Model. We derived different formulas for the probabilities of three types of fragments. For each mosaic fragment, we use the lengths of three sequences: L, the polymorphic stretch, f_1 and f_2, the flanking sequences (Fig. 2*A*). The imported fragment must have spanned length L but could have been longer because of the import of nucleotides that were identical in donor and recipient. The model considers all possible imports extending to the left (d_1) and right (d_2) of the polymorphic stretch (Fig. 2*A*). The probability that the sequenced flanking stretches {min(d_1, f_1) [minimum of d_I and f_I]; min(d_2, f_2)} are identical in donor, and recipient was estimated by the $p_{\text{ident}}[n]$ function, which was derived from experimental data (see below). The summed probability of generating the observed distances for all possible import sizes is

$$r\sum_{d_1=0}^{\infty}\sum_{d_2=0}^{\infty}\rho[d_1+L+d_2]p_{\text{ident}}[\min(d_1,f_1)]p_{\text{ident}}[\min(d_2,f_2)] \qquad \textbf{[1]}$$

where r is the recombination rate and $\rho[n]$ is a size distribution yielding the probability that a recombination event is of length n.

Fragments with single polymorphisms can arise when L is 1 bp or by mutation. In this case, the total probability is Eq. **1** plus the mutation rate, μ.

The third formula was needed to calculate the probability of fragments that are identical between paired isolates. Under the assumption that multiple recombinational/mutational events per gene fragment are rare, the probability of identical fragments

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: SNP, single nucleotide polymorphism.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AJ418065–AJ418366).

[¶] To whom reprint requests should be addressed. E-mail: ssuerbaum@hygiene.uni-wuerzburg.de.

www.pnas.org/cgi/doi/10.1073/pnas.251396098

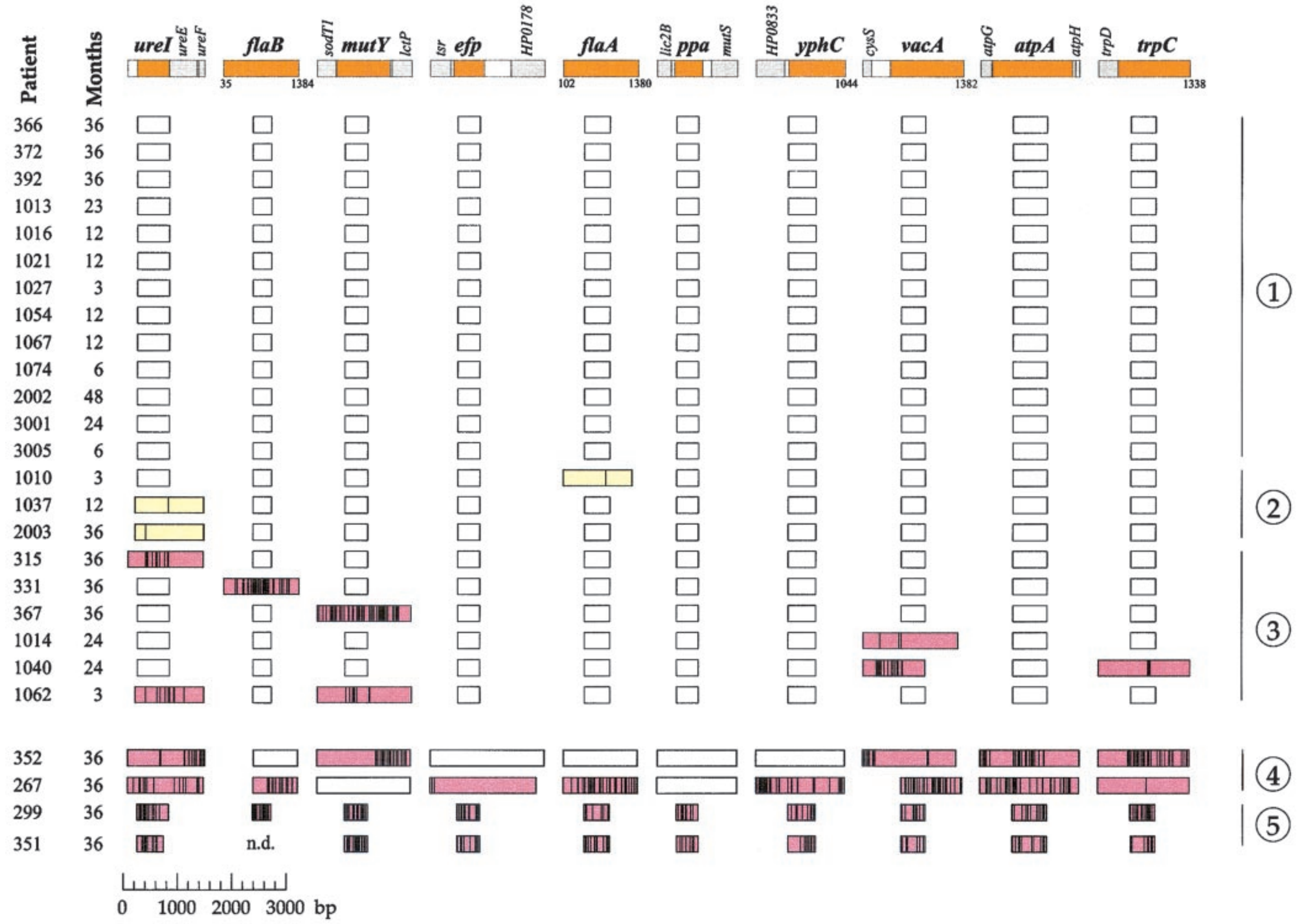

Fig. 1. Sequence comparisons of ten gene fragments from 26 pairs of isolates of *H. pylori*. The ten loci are indicated by tiny maps at the top of the figure (colors: core fragments, tan; flanking genes, gray; noncoding regions, white). The sources of the isolates are indicated by patient codes at the left (Colombia, three digits; Louisiana, four digits). The boxes indicate the lengths of the sequenced fragments, distinguished by color (white, identical sequences; yellow, SNPs; pink, clustered polymorphisms). Vertical lines within the boxes indicate the positions of sequence polymorphisms. The patients are separated into five groups (right) on the basis of all genetic changes between the paired isolates (1, no polymorphisms; 2, only SNPs; 3, clustered polymorphisms; 4, related isolates with numerous polymorphic fragments; 5, unrelated isolates). Mathematical analysis was only performed with data from patients in groups 1–3.

was estimated by subtracting from 1.0 the probability of observing mutational or recombinational events:

$$1.0 - F\mu - r\sum_{x=-\infty}^{F-1}\sum_{y=x+1}^{\infty}\rho[y-x+1]\,(1.0 - p_{\mathrm{ident}}[\min(y,F) - \min(-x,1)+1]) \qquad [2]$$

where F is the size of the sequenced fragment and x and y are the bounds of possible imported fragments (Fig. 2*B*).

$p_{\mathrm{ident}}[n]$ is the proportion of runs of identity between paired sequences that contain at least n nucleotides. Virtual sequences were constructed for each initial isolate consisting of all sequenced fragments joined end to end. $p_{\mathrm{ident}}[n]$ was estimated within and between geographic areas by pairwise comparisons from the first to the last polymorphic site within these virtual sequences. For mosaic fragments, runs of identity were tabulated within each polymorphic stretch between paired isolates.

The size distribution ρ model was implemented by using the exponential distribution.

$$\rho[n] = \frac{1}{\lambda}e^{-\frac{n}{\lambda}} \qquad [3]$$

where λ is the mean recombination size. The exponential distribution is very similar to the geometric size distribution that has been used to model tract lengths for gene conversion in *Drosophila melanogaster* (8). The fit did not improve with a generalized gamma distribution, which contains an extra parameter.

Bayesian Parameter Estimation. The total log likelihood for all data depends on the three parameters r, μ, and λ. The likelihood for individual combinations of these parameters was calculated by summing log likelihoods from the appropriate formula for each fragment. To facilitate numerical calculation, we assumed a maximum imported size of 20,000 bp. Median estimates and credibility regions were obtained from the likelihoods by using the Metropolis algorithm (9). The Metropolis algorithm wanders stochastically through parameter space, preferentially drifting toward combinations of parameters with higher probabilities. Probability calculations depend on priors that set a range of plausible parameter values. Because only 10^5 paired nucleotides had been sequenced, mutations would not have been detected at a frequency below 10^{-5}. Discrete values of $\log_{10}(\mu)$ between -7 and -3 were used as a uniform prior such that half the weight is on detectable frequencies. Continuous values between -7 and

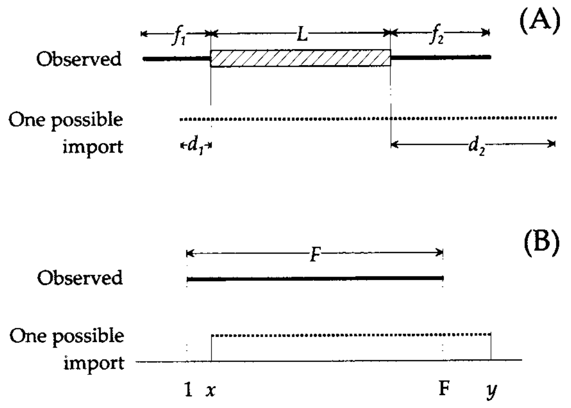

Fig. 2. Sequence lengths used in the model as described in *Materials and Methods*.

−3 were used as a uniform prior on $\log_{10}(r)$. A value of −7 would result in a 2% probability that one or more import ends had been observed while a value of −3 would result in over 100 ends, both of which are outside the range of the observed data. Continuous values of 1 to 4 were used as a uniform prior on $\log_{10}(\lambda)$, which approximates the range of imported DNA in different organisms. The posterior was calculated as the amount of time spent by the Metropolis algorithm in each part of parameter space. Marginal likelihoods (posterior/prior) were calculated for each discrete value of μ. The marginal likelihood for each combination of r and λ was calculated as the sum of the likelihoods for each value of μ weighted by its posterior. Three repeated runs of 100,000 iterations each yielded values of r and λ that differed by less than 3%. Of these estimates, only that for μ is sensitive to the priors for the other parameters. Forced lower priors on r would have resulted in slightly higher estimates of μ. The program implementing this model is available on request.

Results

Sequence Differences in Sequential Isolates. The genetic relationships of sequential isolates of *H. pylori* were investigated for 16 patients in New Orleans, LA (6) and 10 patients in Narino, Colombia (7). For each isolate, a total of 4,658 bp was sequenced from ten "core fragments" at unlinked chromosomal loci that encode seven housekeeping enzymes and three virulence-associated proteins. The same loci have also been sequenced for other isolates from diverse global sources (3, 4). For each core fragment that differed within a pair of isolates, ≈1 kb of flanking DNA was also sequenced to increase the reliability of the estimates of import lengths.

Of the pairs of isolates, 24 of 26 are closely related because large sequence stretches were identical in each pair (Fig. 1, parts 1–4). Indistinguishable whole-genome restriction enzyme patterns for 14 pairs from New Orleans (6) provide independent support for the relatedness within these pairs. The two remaining pairs from New Orleans (patients 1014 and 1040) must also contain closely related strains because of extensive sequence identities, even though they yielded distinct restriction enzyme patterns (6). In contrast, two pairs from Colombia contain genuinely unrelated isolates because their sequences differed at all 10 loci (Fig. 1, part 5); they were not investigated further.

Thirteen pairs of isolates (Fig. 1, part 1) contained no sequence differences. Three pairs of isolates contained only single nucleotide polymorphisms (SNPs) that might reflect mutations or short imports (Fig. 1, part 2). Six pairs of isolates differed by multiple nucleotide exchanges at one or two of the gene fragments and are likely to reflect import of sequences from other bacteria during mixed colonization (Fig. 1, part 3). This set of 22 pairs of sequences comprises the data set that was used for calculations of recombination rate, import size, and mutation rate.

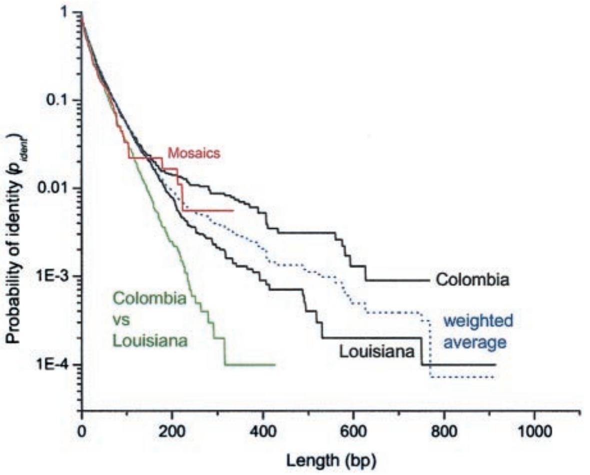

Fig. 3. Probability of identity (p_{ident}) between sequences versus length. The black curves are pairwise comparisons between the initial isolates within Colombia or Louisiana. The weighted average of these curves is shown in blue. The green curve is from pairwise comparisons between Louisiana and Colombia. Data from within polymorphic stretches from paired isolates (Fig. 1, part 3) are in red.

Two other pairs (Fig. 1, part 4) contain related isolates, but so many loci were polymorphic that some polymorphic stretches may reflect more than one import. This possibility is supported by unusually long stretches of sequence identity within their polymorphic stretches (data not shown). The data from these two pairs could lead to biased estimates of recombination size and they were excluded from mathematical analysis.

Length of Identical Sequences Between Unrelated Isolates. A model was devised that calculates the likelihood of the data based on the assumption that each mosaic fragment results from a single event. The likelihood depends on various parameters, including the probability that imported mosaics include flanking nucleotide stretches with identical sequences. This probability, p_{ident}, reflects the diversity of the gene pool of imported sequences. Within-population estimates of p_{ident} based on pairwise comparisons between patients in Colombia and Louisiana were almost indistinguishable up to 200 bp (Fig. 3, Colombia and Louisiana). Between-population estimates were consistently lower (Fig. 3, Columbia vs. Louisiana) than the within-population estimates, indicating that the gene pools differ between these two areas.

p_{ident} from the imported mosaics (Fig. 3, Mosaics) resembled the between population estimate up to lengths of 100 bp and the within population estimates between 100 and 300 bp. This estimate of p_{ident} is not suitable for calculations because it does not extend beyond 300 bp due to a limited sample size; however, it does confirm that the mosaic sequences were indeed imported from *H. pylori* that are similar to the populations sampled here, and justifies the use of population-based p_{ident} estimates. The parameter estimates based on within- and between-population functions of p_{ident} differed by less than 5%. The following parameter estimates were obtained by using an average within population estimate, weighted by the number of pairwise comparisons from each area (Fig. 3, weighted average).

Parameter Estimates. The model calculates the likelihood of the sequence changes and identities in each fragment by summing the probabilities of all possible events that could result in these combinations. The total likelihood for all data depends on the

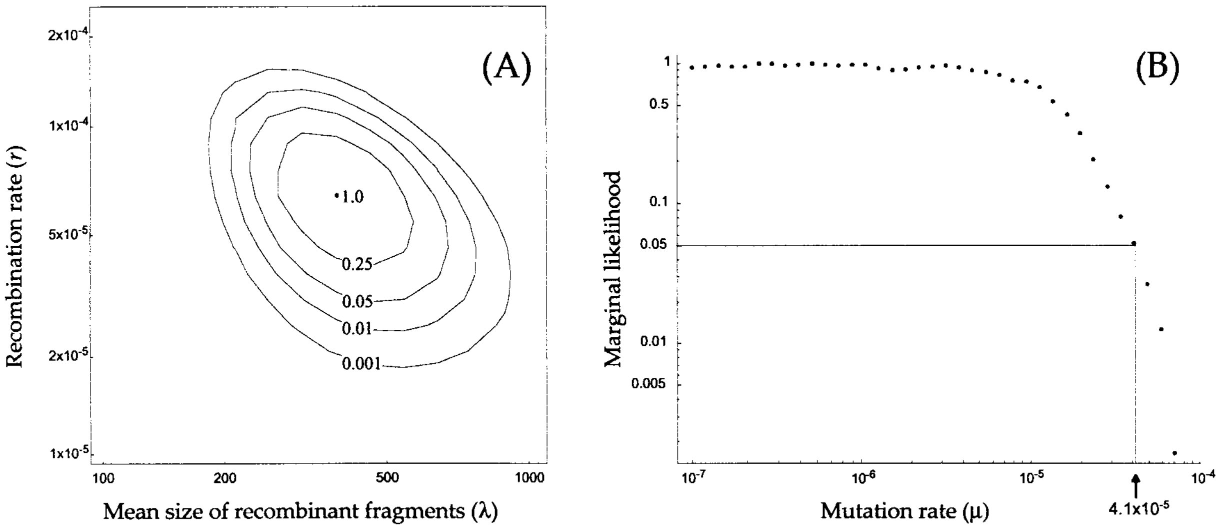

Fig. 4. Parameters estimated by the model for the data in Fig. 1, parts 1–3. (*A*) Contour plot of marginal likelihoods (posterior/prior) of recombination parameters. (*B*) Marginal likelihood of mutation rates. The arrow indicates a maximum below which different mutation rates were not distinguished at the 5% level.

mean recombination size λ, the recombination rate r, and the mutation rate μ. The most probable estimates of r and λ fell into a reasonably narrow range (Fig. 4*A*), despite the presence of only eight mosaics in the data set (Fig. 1, part 3). Analysis of simulated data (see the supporting information, which is published on the PNAS web site, www.pnas.org) showed that the model can accurately estimate mean recombination size when several of the sequences contain at least one end of the imported DNA fragment. According to the p_{ident} distribution for *H. pylori*, the end of an import has probably been reached when it is flanked by identical sequences of over 200 bp. This situation applies to all but one of the sixteen ends in Fig. 1, part 3.

The estimated value of mean recombination fragment size, λ, is 417 bp (95% credibility region of 259–732 bp). This number is considerably lower than estimates of recombination size after transformation or transduction in other bacteria and is comparable to estimates for gene conversion in *D. melanogaster* (Table 1).

The recombination rate, r, is the rate with which recombination events start at any particular nucleotide. The estimated value of r (6.9×10^{-5}; 95% credibility region 3.5×10^{-5} to 1.2×10^{-4}) indicates that each pair of sequential isolates differs on average by 114 (58–200) recombination events. Based on the estimates of λ and r, 2.9% (1.5–4.9%) of the genome or a total of 48,000 nucleotides differ between each sequential pair of isolates because of imported DNA.

The mutation rate, μ, was estimated as being at most 4.1×10^{-5} (Fig. 4*B*), resulting in an r/μ ratio of at least 1. This is only a maximal estimate of μ (and a minimal estimate of r/μ) because the three SNPs that were found might have resulted from recombination rather than mutation according to the model parameters. All three SNPs led to amino acid changes (nonsynonymous exchanges).

r/μ (0.02 [$r = 7 \times 10^{-12}$; $\mu = 3 \times 10^{-10}$]) for neutral genes from *Escherichia coli* (16), which possesses strong clonal population structure, is at least 50-fold lower than our estimate. r/μ estimates that are only slightly lower than our minimal ratio have been calculated for a porin gene under immune selection in panmictic *Neisseria gonorrhoeae* (ref. 17; range 0.1 to 1.4 in different populations) and for humans (refs. 18 and 19; ≈0.5 [$1.3 \times 10^{-8}/2.5 \times 10^{-8}$]).

Minimal Age of *H. pylori*. Over long time periods, nonsynonymous mutations are removed by selection and synonymous mutations, which are more neutral, contribute most to divergence between strains. In *H. pylori*, synonymous sequence polymorphisms in housekeeping genes are four times as frequent between random pairs of isolates as are nonsynonymous differences (data not shown).

A maximal synonymous molecular clock rate (max μ_S) can be calculated for data from paired isolates according to

$$\max\mu_S = \frac{m}{\sum_{i=1}^{no.\,fragments} n_i t_i} \qquad \textbf{[4]}$$

Table 1. Average sizes (bp) of recombined fragments in different organisms

Species	Mean	Median	Source of data	Citation
Drosophila subobscura	122		Population-based, gene conversion	10
D. melanogaster	352		Laboratory, gene conversion	8
H. pylori	417	290	Sequential isolates, transformation	This paper
Streptococcus pneumoniae		2,000	Laboratory, transformation	11
N. meningitidis		5,100	Population-based, transformation	12
S. pneumoniae	6,000		Population-based, transformation	13
Bacillus subtilis		10,000	Laboratory, transformation	14
E. coli	14,000		Laboratory, transduction	15

where n_i is the number of potential synonymous sites in each fragment, t_i is the time between isolation, and m is the number of observed synonymous mutations. Excluding fragments with polymorphic sequence stretches, the data set in Fig. 1 (parts 1–4) contains 22,950 identical synonymous sites between paired isolates taken on average 1.8 years apart, which corresponds to 42,608 synonymous bp years. As described elsewhere (20, 21), the maximal mutation rate can be estimated from zero observed mutations according to the Poisson distribution of e^{-m} by substituting m with 2.996 (95% confidence limit) or 0.693 (50% confidence limit). These substitutions yield corresponding maximal synonymous clock rates of 7×10^{-5} and 1.6×10^{-5}, respectively.

max μ_S can be used to estimate the minimal age of *H. pylori*. To this end, we calculated mean D_S (average pairwise difference at synonymous sites) from sequences of the core fragments from a globally representative collection [220–235 isolates except for *flaA* (72) and *flaB* (103); data not shown]. After Jukes–Cantor correction, the mean D_S for *H. pylori* is 0.182. After division by the mutation rate, this yields a minimal estimate of 11,000 years (50% confidence limit) or 2,500 years (95% limit) since the existence of the last common ancestor of *H. pylori*.

The synonymous clock rate per nucleotide differs between bacterial species (22). The maximal clock rate estimated here is roughly four orders of magnitude faster than that of *E. coli* or *Buchnera* (22) (or mammals). If the true clock rate were as low as the *E. coli* rate, the diversity in *H. pylori* would indicate that the last common ancestor of all *H. pylori* existed 40 million years ago. This seems unlikely because *H. pylori* should then be isolated from numerous different species of mammals, which is not the case.

Discussion

Family studies are an invaluable tool for discerning patterns of genetic linkage among humans and other eukaryotes. The analysis of sequential bacterial isolates is conceptually similar to family studies and provides an exciting approach for determining basic evolutionary parameters. Bacterial adaptation to the human host can result in amino acid changes in exposed outer membrane proteins (23) and genomic rearrangements (24), but such studies had not yet been performed with selectively neutral genes. This study presents extensive sequence data on numerous pairs of sequential isolates and demonstrates that such data can be used to determine mutation rate, recombination size, and recombination rate.

The data were obtained with pairs of bacteria isolated from adults (mean age of 49 years). The time interval between each pair of isolates and their common ancestor is at least the interval between isolation dates (Fig. 1; average of 1.8 years). This time would be greater if the adults had been colonized with both strains over a longer time period, such as since childhood, or if the strains had already coexisted in the source of infection. Thus, our recombination and mutation clock rates are both maximal estimates.

Recombination. On average, pairs of bacteria differed by ≈100 DNA imports, corresponding to three percent of the genome or 50 kb. By further extrapolation from the average time of 1.8 years between isolation, half of the genome would have been replaced by import within 41 years (1.8 × 22.5; calculated by solving 0.97^x = proportion of genome that is unrecombined). Even if the pairs of isolates were derived from a common ancestor that last existed in early childhood, half of the genome would have been replaced within 2,200 years, a surprisingly short time interval. By comparison, 10–100 million years were needed to replace 60% of the *E. coli* genome (25).

Recombination in other bacteria is less frequent. For example, excluding the import of sequences flanking the *tbpB* gene due to selection by the immune system, only one case of import was detected among three gene fragments from 200 isolates of *Neisseria meningitidis* during several years of epidemic and endemic disease (12). Similarly, only three recombination events over 100 kb distinguish five isolates of *E. coli* that are thought to have diverged within the last 2,400 years (26). No recombinations (or point mutations) were detected in six housekeeping gene fragments among 36 isolates of *Yersinia pestis* that have diverged in the last 1,500 years (21). Thus, the recombination frequency within *H. pylori* is extraordinarily high!

Import Size. The mean size of imported fragments in *H. pylori* is unusually small for bacteria and is comparable to the size of gene conversion in *Drosophila* (Table 1). The unusually small size of imported fragments might reflect digestion of naked DNA in the gastric mucosal environment populated by *H. pylori*, the presence of multiple restriction endonucleases in the *H. pylori* genome (27), or still other factors. Combined with the high frequency of import, small imported fragments suggest that *H. pylori* may be lacking mechanisms that restrict DNA import from unrelated organisms. Specific sequences, such as DNA uptake sequences in *Hemophilus influenzae* and the neisseriae (28) or chi-sequences in diverse bacteria (29), ensure that import of DNA is more efficient from related than from unrelated organisms. Such sequences are probably lacking in *H. pylori* (30) and would also not be expected to occur routinely within the short coding fragments that were imported. In enteric bacteria, mismatch repair provides a barrier against import from unrelated organisms (31), but a complete mismatch repair system has also not been identified in *H. pylori* (32).

Possibly *H. pylori* does not need to defend against import of DNA from unrelated organisms. It lives under a protective mucus layer (33) in an isolated, sterile environment without microbial competitors. It has been suggested that DNA import is important for adaptation to the individual host (34) and indeed one import detected here resulted in an inactive *vacA* gene, due to an imported stop codon (data not shown). However, extensive sequence based evidence for the importance of import to host adaptation is still lacking (35). Furthermore, imports were not concentrated on any particular gene, unlike the situation with *N. meningitidis* (12). Thus, the biological significance of frequent import of short fragments remains uncertain and may largely represent neutral events.

Age of *H. pylori*. No synonymous mutations were detected during a total of 42,608 synonymous bp years, leading to a maximal synonymous clock rate of 2–7 × 10^{-5} and a corresponding minimal age since the last common ancestor of 2,500–11,000 years. Refined estimates could be obtained through sequencing additional fragments from the same 22 pairs of isolates and/or by comparison of paired isolates from families that have been separated for longer times.

H. pylori has been isolated from humans across the globe and sequence differences between different continents indicate that these bacterial populations have been separated for millenia (4). The length of time with which *H. pylori* has been associated with humans is interesting in the context of important human milestones, such as the ≈13,000 years of agriculture since the end of the last ice age and the ≈50,000 years since global colonization by anatomically modern man (36). The current minimal age estimate of 2,500–11,000 years needs to be refined by one to two orders of magnitude for comparison with these milestones.

Summary. The results presented here provide a paradigm for estimating basic evolutionary parameters of bacteria based on the use of sequential isolates. A method is described that can reliably estimate recombination rates and mean imported fragment size based on limited numbers of events, as long as the ends of the imports are included within the sequenced fragments. The method also has the potential to accurately determine mutation clock rates with larger data sets than were used in this analysis.

The data presented here provide direct evidence that the panmictic population structure of *H. pylori* is caused by very frequent recombination during mixed colonization by unrelated strains. *H. pylori* is characterized by the highly unusual combination of high import frequency and low import size. Recombination is so frequent that appreciable fractions of the entire genome are exchanged during the colonization of a single human, resulting in a highly flexible genome content and frequent shuffling of sequence polymorphisms throughout the local gene pool.

We appreciate helpful comments on the manuscript by Christine Josenhans, Giovanna Morelli, and Helga Gressmann, and expert technical assistance by Susanne Friedrich, Anne Wirsing, and Elena Katzowitsch. This work was supported by grants SU 133/3-2 and SFB 479/A5 from the Deutsche Forschungsgemeinschaft, and by the EBP network (EU contract QLK2-CT-2000-01536).

1. Feldman, R. A. (2001) in *Helicobacter pylori: Molecular and Cellular Biology*, eds. Achtman, M. & Suerbaum, S. (Horizon Scientific Press, Norfolk, England), pp. 29–51.
2. Achtman, M. (2001) in *Helicobacter pylori: Molecular and Cellular Biology*, eds. Achtman, M. & Suerbaum, S. (Horizon Scientific Press, Wymondham, U.K.), pp. 311–321.
3. Suerbaum, S., Maynard Smith, J., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12619–12624.
4. Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z.-J., Suerbaum, S., Thompson, S., van der Ende, A. & van Doorn, L. J. (1999) *Mol. Microbiol.* **32,** 459–470.
5. Kersulyte, D., Chalkauskas, H. & Berg, D. E. (1999) *Mol. Microbiol.* **31,** 31–43.
6. Taylor, N. S., Fox, J. G., Akopyants, N. S., Berg, D. E., Thompson, N., Shames, B., Yan, L., Fontham, E., Janney, F., Hunter, F. M., *et al.* (1995) *J. Clin. Microbiol.* **33,** 918–923.
7. Correa, P., Fontham, E. T., Bravo, J. C., Bravo, L. E., Ruiz, B., Zarama, G., Realpe, J. L., Malcom, G. T., Li, D., Johnson, W. D., *et al.* (2000) *J. Natl. Cancer Inst.* **92,** 1881–1888.
8. Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H. & Chovnick, A. (1994) *Genetics* **137,** 1019–1026.
9. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
10. Betran, E., Rozas, J., Navarro, A. & Barbadilla, A. (1997) *Genetics* **146,** 89–99.
11. Guild, W. R., Cato, A., Jr., & Lacks, S. (1968) *Cold Spring Harbor Symp. Quant. Biol.* **33,** 643–645.
12. Linz, B., Schenker, M., Zhu, P. & Achtman, M. (2000) *Mol. Microbiol.* **36,** 1049–1058.
13. Enright, M. C. & Spratt, B. G. (1999) *Mol. Biol. Evol.* **16,** 1687–1695.
14. Fornili, S. L. & Fox, M. S. (1977) *J. Mol. Biol.* **113,** 181–191.
15. McKane, M. & Milkman, R. (1995) *Genetics* **139,** 35–43.
16. Milkman, R. & Bridges, M. M. (1990) *Genetics* **126,** 505–517.
17. Posada, D., Crandall, K. A., Nguyen, M., Demma, J. C. & Viscidi, R. P. (2000) *Mol. Biol. Evol.* **17,** 423–436.
18. Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A. J., Deloukas, P., Olsen, A., Doggett, N. A., Ghebranious, N., Broman, K. W., *et al.* (2001) *Nature (London)* **409,** 951–953.
19. Nachman, M. W. & Crowell, S. L. (2000) *Genetics* **156,** 297–304.
20. Rich, S. M., Licht, M. C., Hudson, R. R. & Ayala, F. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4425–4430.
21. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. & Carniel, E. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14043–14048.
22. Ochman, H., Elwyn, S. & Moran, N. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12638–12643.
23. Groeneveld, K., van Alphen, L., Voorter, C., Eijk, P. P., Jansen, H. M. & Zanen, H. C. (1989) *Infect. Immun.* **57,** 3038–3044.
24. Römling, U., Schmidt, K. D. & Tümmler, B. (1997) *FEMS Microbiol. Lett.* **150,** 149–156.
25. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44,** 383–397.
26. Guttman, D. S. & Dykhuizen, D. E. (1994) *Science* **266,** 1380–1383.
27. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., *et al.* (1997) *Nature (London)* **388,** 539–547.
28. Goodman, S. D. & Scocca, J. J. (1991) *J. Bacteriol.* **173,** 5921–5923.
29. El Karoui, M., Biaudet, V., Schbath, S. & Gruss, A. (1999) *Res. Microbiol.* **150,** 579–587.
30. Saunders, N. J., Peden, J. F. & Moxon, E. R. (1999) *Microbiology* **145,** 3523–3528.
31. Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., *et al.* (2000) *Cell* **103,** 711–721.
32. Wang, G., Humayun, M. Z. & Taylor, D. E. (1999) *Trends Microbiol.* **7,** 488–493.
33. Schreiber, S., Stuben, M., Josenhans, C., Scheid, P. & Suerbaum, S. (1999) *Infect. Immun.* **67,** 5151–5156.
34. Montecucco, C. & Rappuoli, R. (2001) *Nat. Rev. Mol. Cell Biol.* **2,** 457–466.
35. Kuipers, E. J., Israel, D. A., Kusters, J. G., Gerrits, M. M., Weel, J., van Der, E. A., Der Hulst, R. W., Wirth, H. P., Hook-Nikanne, J., Thompson, S. A., *et al.* (2000) *J. Infect. Dis.* **181,** 273–282.
36. Diamond, J. (1997) *Guns, Germs and Steel* (Jonathan Cape, London).