Genetic variation in Helicobacter pylori
Pan, Z.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions

*Molecular Microbiology* 1999; 32: 459-470
Chapter 6

Recombination and Chromosome Mapping

Heterocryptic Behavior from Different Geographical Regions
Recombination and clonal groupings within Helicobacter pylori from different geographical regions

Mark Achtman,1 Takeshi Azuma,2 Douglas E. Berg,3 Yoshiyuki Ito,2 Giovanni Morelli,1 Zhi-Jun Pan,3 Sebastian Suerbaum,4 Stuart A. Thompson,5 Arie van der Ende6 and Leen-Jan van Doorn7

1Max-Planck Institut für molekulare Genetik, Ihnestr. 73, 14195 Berlin, Germany.
2Second Department of Internal Medicine, Fukui Medical University, Fukui 910-1193, Japan.
3Departments of Molecular Microbiology and Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA.
4Department of Medical Microbiology, Ruhr-Universität Bochum, 44780 Bochum, Germany.
5Vanderbilt University School of Medicine, Nashville, TN 37232, USA.
6Department of Medical Microbiology, Academic Medical Center, 1105 AZ Amsterdam, The Netherlands.
7Delft Diagnostic Laboratory, 2625 AD Delft, The Netherlands.

Summary
A collection of 20 strains of Helicobacter pylori from several regions of the world was studied to better understand the population genetic structure and diversity of this species. Sequences of fragments from seven housekeeping genes (atpA, efp, mutY, ppa, trpC, ureI, yphC) and two virulence-associated genes (cagA, vacA) showed high levels of synonymous sequence variation (mean percentage $K_a$ of 10–27%) and lower levels of non-synonymous variation (mean percentage $K_a$ of 0.2–5.6%). Cluster analysis of pairwise differences between alleles revealed the existence of two weakly clonal groupings, which included half of the strains investigated. All six strains isolated from Japanese and coastal Chinese were assigned to the ‘Asian’ clonal grouping, probably reflecting descent from a distinct common ancestor. The clonal groupings were not totally uniform; recombination, as measured by the homoplasy test and compatibility matrices, was extremely common within all genes tested, except cagA. The fact that clonal descent could still be discerned despite such frequent recombination possibly reflects founder effects and geographical separation and/or selection for particular alleles of these genes.

Introduction
Helicobacter pylori is a Gram-negative bacterium that infects the gastric mucosa of more than half of all humans. It is a major cause of peptic ulcers and an early risk factor for gastric cancer (Blaser, 1997). H. pylori is also genetically more diverse than most bacterial species (Go et al., 1996). The DNA fingerprint pattern and sequences of various genes are almost always different between independent pairs of isolates (Majewski and Goodwin, 1988; Akopyanz et al., 1992a,b; Kansau et al., 1996), and a comparison of the genomes of two strains has shown that 7% of the genes are specific to each strain (Alm et al., 1999). However, bacteria isolated from the members of one family can be indistinguishable, or show only limited variation, which implies intrafamilial transmission (Bamford et al., 1993; van der Ende et al., 1996; Suerbaum et al., 1998; Kersulyte et al., 1999).

The limited data available to date on the population genetics of this organism have led to somewhat contradictory interpretations. For example, multilocus enzyme electrophoresis of six loci indicated no linkage disequilibrium between the different loci (Go et al., 1996), and restriction fragment length polymorphism (RFLP) analysis indicated that its population genetic structure is panmictic (Salaun et al., 1998). Similarly, analyses of sequences from fragments of the vacA, flaA and flaB genes indicated ‘free recombination’ among isolates from Canada and Germany (Suerbaum et al., 1998), i.e. no clonal associations were detected, even among pairs of polymorphic sites within single genes. In agreement, multiple recombinant types were found among strains isolated from one individual (Kersulyte et al., 1999).

The virulence-associated vacA gene encodes a vacuolating cytotoxin (Cover, 1996; Atherton, 1998). Particular sequence motifs in the signal sequence (‘s’) and middle (‘m’) regions of the vacA gene differ between H. pylori...
isolated from East Asians and Europeans (Atherton et al., 1995; Itō et al., 1996; Pan et al., 1998; van Doorn et al., 1998a), suggesting that recombination is rare between bacteria from different continents or that particular alleles are selected for in certain populations. However, a different segment of the vacA gene was found to have recombined freely in bacteria isolated from Canada and South Africa (Suerbaum et al., 1998).

The virulence-associated cagA gene encodes an immunodominant protein of unknown function (Covacci et al., 1993; Tummuru et al., 1993; Kuipers et al., 1995). cagA is located at one end of a 40 kb pathogenicity island (Censini et al., 1996; Akopyants et al., 1998) that is lacking from many putatively avirulent strains from the USA and Europe and is altered in other strains (Jenks et al., 1998). Similarly to vacA, the population of cagA alleles differs between strains isolated from East Asians and strains from other ethnic groups (Miehlke et al., 1996; Pan et al., 1997; von der Ende et al., 1998; van Doorn et al., 1999). Thus, recombination might be frequent among strains isolated from individual ethnic groups but rare between strains from geographically separated ethnic groups. However, the same Chinese and Dutch strains that had exhibited geographical partitioning of cagA sequences did not show ethnic associations for sequences of the glmM housekeeping gene (van der Ende et al., 1998). And the same polymorphisms were found in the vacA, flaA and flaB genes from South Africa and Germany or Canada (Suerbaum et al., 1998).

Recently, sequences from multiple housekeeping genes (multilocus sequence typing, MLST) have been used to define phylogenetic relationships within the weakly clonal species Neisseria meningitidis (Maiden et al., 1998) and Streptococcus pneumoniae (Enright and Spratt, 1998). Most members of each clonal grouping within these species are uniform for multiple housekeeping genes; recombinant alleles are the exception rather than the rule. It seemed possible, especially considering the geographical distributions of cagA and vacA described above, that MLST would also allow the recognition of distinct clonal groupings within H. pylori. The conclusion that H. pylori is rarely recombining was based on vacA, flaA and flaB sequences from strains isolated in Canada, Germany and South Africa; it seemed possible that analysis of a more global collection might reveal geographical associations that were lacking in the earlier sample. Furthermore, alleles of cytoplasmic housekeeping genes are more likely to be selectively neutral, and therefore more uniform, than is the case for genes for which there might be selection for diversity as a result of association with virulence (vacA) or motility (flaA, flaB).

In this report, we analyse the sequence variability of both housekeeping and virulence-associated genes from H. pylori strains isolated in several continents and show that weakly clonal groupings can be identified in H. pylori despite a rich history of interstrain recombination.

Results

A global strain collection

We combined strains of H. pylori that have been used for various laboratory studies with additional strains from diverse geographical locations to assemble a collection of 19 strains that may represent much of the global population

<table>
<thead>
<tr>
<th>Strain</th>
<th>Country</th>
<th>Clonal group</th>
<th>cagA</th>
<th>vacA</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCTC11638</td>
<td>Australia</td>
<td>Australia</td>
<td>m1</td>
<td>m2a</td>
<td>AvdE</td>
<td>van der Ende et al. (1998)</td>
</tr>
<tr>
<td>25</td>
<td>Holland</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BO265</td>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26095</td>
<td>UK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-1</td>
<td>Lithuania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Strains of H. pylori whose sequences were analysed.

Strain NCTC11638 has also been referred to as BO265, J166 as 95-58 and HP1 as Peru2-11. 

a. Source, first letters of author's name from whom the strain or its DNA may be obtained.
of this species (Table 1). Corresponding sequences from strain 26695, whose entire genome has been sequenced (Tomb et al., 1997), were also included in the analysis. [The genome sequence of strain J99 (Alm et al., 1999) was not included because it was published after completion of the analyses presented here.] The strains were tested by hybridization with probes specific for cagA and for the s (signal peptide) and m (internal) region alleles of vacA (Atherton et al., 1995; van Doorn et al., 1998a,b). All strains hybridized with the cagA probe except HP1, in which much of the cag pathogenicity island including part of cagA is deleted (D. Kersulyte and D. E. Berg, unpublished). All strains except one (CC28) yielded unique hybridization patterns with probes for the s and m regions of vacA. The results (Table 1) indicated that the vacA s1c allele was present exclusively in the six strains isolated from Asians, in agreement with published data (van Doorn et al., 1998a), whereas the other vacA alleles were not associated exclusively with any geographical or ethnic source.

**Sequences**

Sequences were obtained from fragments of seven housekeeping genes scattered around the chromosome, from a fragment of vacA and from both N-terminal (termed cagAN) and internal (cagAC) fragments of cagA (see Experimental procedures) (Table 2). Together with the published sequence from strain 26695, this resulted in a total of 20 sequences per gene fragment. One other strain (HP1) also lacked much of the cagA gene, as indicated above. Accordingly, our analyses were restricted to 19 sequences for cagAN and cagAC.

**Sequence variation**

After alignment, none of the sequences contained gaps or insertions. The average GC content of the entire *H. pylori* genome is 39% (Tomb et al., 1997), but large regions can contain as little as 33% or as much as 43%. Most of the gene fragments analysed here possessed GC contents of ≈38%, but three possessed GC contents of ≈45% (atpA, efp, ureI; Table 3).

Only a few amino acid changes (non-synonymous sequence variation) were found in most gene fragments (mean percentage $K_a$ 0.2–3.1%), but the mean percentage $K_a$ of both cagAN and cagAC was 5.5%. All sequences showed high levels of synonymous sequence variation, with mean percentage $K_s$ values of 12–27%; individual

---

**Table 2. Gene fragments that were sequenced, ordered by position in the genome of strain 26695.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number</th>
<th>Position (in gene)</th>
<th>Length (bp)</th>
<th>Gene product or inferred function</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>uroF</td>
<td>HP0071</td>
<td>74,750 (1)</td>
<td>685</td>
<td>Urease accessory protein</td>
<td>AJ239701-19</td>
</tr>
<tr>
<td>mutY</td>
<td>HP0142</td>
<td>154,138 (409)</td>
<td>420</td>
<td>A/G-specific adenosine deaminase</td>
<td>AJ239645-63</td>
</tr>
<tr>
<td>efp</td>
<td>HP0177</td>
<td>183,801 (76)</td>
<td>410</td>
<td>Elongation factor EF-P</td>
<td>AJ239507-22</td>
</tr>
<tr>
<td>cagA</td>
<td>HP547</td>
<td>579,972 (52)</td>
<td>413</td>
<td>Cytobolin-associated gene A</td>
<td>AJ239833-700</td>
</tr>
<tr>
<td>cagAN</td>
<td>HP0620</td>
<td>666,708 (22)</td>
<td>243</td>
<td>Central vacA fragment</td>
<td>AJ239720-37</td>
</tr>
<tr>
<td>vacA</td>
<td>HP0834</td>
<td>885,070 (5)</td>
<td>610</td>
<td>Inorganic polyphosphatase</td>
<td>AJ239628-44</td>
</tr>
<tr>
<td>cagAC</td>
<td>HP0887</td>
<td>936,619 (205)</td>
<td>444</td>
<td>Glycine tRNA synthetase</td>
<td>AJ239664-82</td>
</tr>
<tr>
<td>ppa</td>
<td>HP1134</td>
<td>1,197,139 (512)</td>
<td>629</td>
<td>Vacuolating cytotoxin</td>
<td>AJ239588-906</td>
</tr>
<tr>
<td>atpA</td>
<td>HP1273</td>
<td>1,354,937 (864)</td>
<td>456</td>
<td>ATP synthase, F1</td>
<td>AJ239569-87</td>
</tr>
</tbody>
</table>

Longer sequences for yphC (1071 bp) and cagAN (up to 484 bp) were trimmed to the lengths shown such that all sequences evaluated were from intragenic regions and did not contain any alignment gaps. Position (in gene), TIGR position of the N-terminal end of the fragment sequenced (position relative to start of gene, which was defined as 1).

---

**Table 3. Properties of gene fragments sequenced.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>%GC Mean percentage $K_s$ (range)</th>
<th>Mean percentage $K_a$ (range)</th>
<th>Ratio $K_a/K_s$</th>
<th>Mean percentage compatibility</th>
<th>$H$ ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>efp</td>
<td>46.0 12.1 ± 3.2 (5.5–19.5)</td>
<td>0.2 ± 0.2 (0–0.6)</td>
<td>49.6</td>
<td>48</td>
<td>0.76</td>
</tr>
<tr>
<td>cagAN</td>
<td>37.0 12.4 ± 6.2 (0–25.6)</td>
<td>5.6 ± 3.8 (0.6–12.4)</td>
<td>2.2</td>
<td>62</td>
<td>0.17</td>
</tr>
<tr>
<td>cagAC</td>
<td>38.0 24.1 ± 18.6 (0–65.5)</td>
<td>5.3 ± 4.2 (0–13.0)</td>
<td>4.5</td>
<td>77</td>
<td>0.15</td>
</tr>
<tr>
<td>atpA</td>
<td>46.0 14.6 ± 5.7 (0–31.6)</td>
<td>0.3 ± 0.3 (0–1.3)</td>
<td>48.7</td>
<td>50</td>
<td>0.60</td>
</tr>
<tr>
<td>mutY</td>
<td>39.0 22.3 ± 7.3 (6.6–41.4)</td>
<td>1.9 ± 0.8 (0.3–3.9)</td>
<td>11.9</td>
<td>54</td>
<td>0.66</td>
</tr>
<tr>
<td>ppa</td>
<td>41.0 10.5 ± 4.0 (0–23.9)</td>
<td>0.5 ± 0.3 (0–1.6)</td>
<td>22.9</td>
<td>54</td>
<td>0.68</td>
</tr>
<tr>
<td>trpC</td>
<td>39.0 26.8 ± 8.9 (6.1–43.6)</td>
<td>3.1 ± 1.2 (0–5.4)</td>
<td>8.5</td>
<td>55</td>
<td>0.46</td>
</tr>
<tr>
<td>ureI</td>
<td>44.0 11.6 ± 4.8 (3.8–26.1)</td>
<td>0.7 ± 0.4 (0–1.9)</td>
<td>16.5</td>
<td>45</td>
<td>0.79</td>
</tr>
<tr>
<td>vacA</td>
<td>43.0 17.8 ± 6.1 (3.2–36.4)</td>
<td>2.8 ± 1.5 (0–7.5)</td>
<td>6.9</td>
<td>52</td>
<td>0.69</td>
</tr>
<tr>
<td>yphC</td>
<td>38.0 17.5 ± 5.5 (5.5–31.3)</td>
<td>1.8 ± 0.8 (4–4.2)</td>
<td>9.7</td>
<td>42</td>
<td>0.67</td>
</tr>
</tbody>
</table>
pairs of sequences differed at as many as 66% of synonymous sites. Because non-synonymous variation was frequent within cagA, the $K_s/K_a$ ratio was considerably lower for both fragments from this gene than from the other genes (9-50). All sequences were unique, except that identical efp sequences were found in strains 96-232 and 92-228 (both from Alaska), and identical ppa sequences were found in strains 96-228 (Alaska) and J165 (USA). These results indicate, in agreement with previous analyses (Garner and Cover, 1995; Kansau et al., 1998; Suerbaum et al., 1998; van der Ende et al., 1998; van Doorn et al., 1998a, b), that identical sequences are rare among H. pylori.

Frequency of recombination

The sets of sequences were tested by the homoplasy test (Maynard Smith and Smith, 1998), which measures the frequency of the same nucleotide changes in different branches of a maximum parsimony phylogenetic tree (homoplasies). Homoplasies are caused by horizontal DNA transfer between unrelated strains (recombination) or by independent mutations. The homoplasy test calculates the $H$ ratio, the frequency of observed synonymous homoplasies relative to the frequencies expected for the observed sequence variation under clonal descent or free recombination. If all descent is clonal (no interstrain recombination, $H$ ratio = 0.0), all homoplasies arise by the coinnoculation accumulation of mutations at the same sites in independent strains. In contrast, under free recombination ($H$ ratio = 1.0), mutations need only arise once and are then acquired by sequences in other genomes via recombination. Sequences from the flaA, flaB and vacA genes from strains isolated in Germany or Canada had yielded $H$ ratios between 0.8 and 0.9 (Suerbaum et al., 1998).

With the current collection of 20 strains, the vacA gene fragment yielded an $H$ ratio of 0.69, only slightly lower than observed previously, indicating frequent recombination. The $H$ ratio for six of the housekeeping genes was between 0.60 and 0.79 (geometric mean of 0.69), indicating similarly high levels of recombination for these genes (Table 3). The seventh housekeeping gene, trpC, yielded an intermediate $H$ ratio of 0.46. In contrast to these results, the two cagA gene fragments yielded $H$ ratios of 0.15–0.17, among the lowest ratios yet observed in any species (Suerbaum et al., 1998). Thus, most genes in H. pylori exhibit free recombination, while recombinants involving at least some alleles of the cagA gene are rare.

The sequences were also analysed by compatibility matrices (Jakobsen and Eastsea, 1996) (Fig. 1). This test scores whether pairs of informative sites are compatible with a maximum parsimony tree (clonal descent), as indicated in Fig. 1 by a white box, or incompatible (recombination), as indicated by a black box. The matrices are largely black, implying repeated recombination events, except for the cagAC gene fragment in which large regions are white, compatible with clonal descent. A numerical measure of the degree of clonal descent is the average frequency with which pairs of sites are compatible, called the mean percentage compatibility. The mean percentage compatibilities of the matrices were between 42% and 62% (Table 3), except for cagAC, which was 77%. These results indicate that interstrain recombination was frequent within all gene sets except cagAC.

The data from both analyses concur that there has been extensive interstrain recombination in the H. pylori chromosome, with the notable exception of the cagAC gene fragment, and intermediate or contradictory values for the trpC and cagAN gene fragments.

Clonal descent

MLST clonal groupings in N. meningitidis and S. pneumoniae have been defined as sets of bacteria sharing a significant number of identical alleles (Enright and Spratt, 1998; Maiden et al., 1998). Identical alleles are so rare in H. pylori that this criterion does not apply to this species. Instead, we examined the average sequence homology for the eight gene fragments for which sequences had been obtained from all 20 strains (thus excluding cagA). Normalized distance matrices of the numbers of nucleotide differences between each of the pairs of sequences were calculated for each gene and averaged to yield a mean distance matrix. An ureA (unweighted pair group mean average) cluster analysis of this matrix showed that the six strains isolated from East Asians clustered together, including strain HP1, which was from an ethnic Japanese living in Peru (Fig. 2). We refer to this group of strains as the 'Asian' clone. Similarly, three other strains (#12, 5596 and CC28; isolated in Guatemala, The Gambia and South Africa respectively) also clustered together and separately from the other strains. These bacteria are referred to as 'clone2' (Fig. 2). Very similar results were obtained using a mean distance matrix from pairwise distances that had not been normalized.

The sequences from each of the 10 gene fragments were also analysed individually by bootstrap tests of neighbour-joining trees and by maximum likelihood cluster analysis. The results showed that, for each gene fragment except ppa and alpA, two to six strains of the Asian clone clustered together with bootstrap values of at least 47% (Table 4). Furthermore, for each of the six Asian strains, four to eight of the 10 gene fragments clustered with those from other strains of the Asian clone. Similarly, except for ureA and ppe, all gene fragments usually clustered together for the strains of clone2 (Table 5). These clusters also included strain J166 for four of the gene fragments (Table 5). This group of four strains was also uniform for
Fig. 1. Compatibility matrices of 10 gene fragments from 20 strains of *H. pylori*. These matrices depict the informative sites within the sequence sets as white boxes when pairs of sites are compatible with a parsimony tree and as black boxes for pairs of sites that are incompatible. Black regions indicate runs of incompatible sites that probably arose by repeated recombination, whereas white regions (in particular in the cagAC gene fragment) indicate that these runs of sites are compatible with clonal descent by the sequential accumulation of mutations.

Fig. 2. An unweighted dendrogram of the mean normalized pairwise differences between 20 alleles for eight gene fragments (excluding cagAN and cagAC). The number of nucleotide differences between each pair of alleles was normalized to a maximal value of 1.0 before averaging over the eight gene fragments. The resulting matrix was used for a cluster analysis using the unweighted pair-group mean average clustering method. Note that strain J166 is not in clone2 according to these results, although many gene fragments from J166 did cluster with clone2 in dendrograms of the individual genes (Table 5).
Table 4. Gene fragments whose sequences cluster significantly within six strains isolated from East Asians.

<table>
<thead>
<tr>
<th>Gene</th>
<th>ureI</th>
<th>mutY</th>
<th>stdp</th>
<th>cagA(N)</th>
<th>cagA(C)</th>
<th>pps</th>
<th>yphC</th>
<th>vacA</th>
<th>stpA</th>
<th>stpC</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position (kb)</td>
<td>75</td>
<td>154</td>
<td>184</td>
<td>580</td>
<td>581</td>
<td>666</td>
<td>886</td>
<td>939</td>
<td>1197</td>
<td>1355</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strains</th>
<th>R28</th>
<th>R29</th>
<th>88-28</th>
<th>97-42</th>
<th>HP1</th>
<th>88-39</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of strains</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Bootstrap (%)</td>
<td>85</td>
<td>95</td>
<td>63</td>
<td>99</td>
<td>100</td>
<td>64</td>
</tr>
</tbody>
</table>

Sequences from Asian strains that clustered in neighbour-joining dendrograms with a percentage bootstrap value of at least 47% are indicated by a '+', and other sequences are indicated by an empty space. ND, not tested because a sequence was not obtained.

The clonal associations seen here would probably not have been considered significant if only single housekeeping genes had been analysed. The dendrograms of the individual genes were quite different (see Fig. 3 for examples), and strains that clustered significantly for individual genes (e.g. the two unlabelled clusters for mutY in Fig. 3) were often unrelated for other genes. The dendrogram of the cag\(A\) gene fragments differed from the dendrograms for other loci in that three distinct branches were observed, one of which included the five cag\(A\) Asian strains and a second contained only strain J166, which was otherwise often associated with clone2 (Fig. 3).

**Discussion**

We studied the sequences of fragments of seven housekeeping and two virulence-associated genes from 20 strains of *H. pylori* isolated from diverse geographical regions in order to assess its population genetic structure and evolutionary history. Two weakly clonal groupings were found, superimposed on a pattern of free recombination, an outcome that provides a basis for further population genetic analyses.

**Clonality within panmictic species**

*H. pylori* does not possess a strong clonal structure (Go et al., 1996; Salaun et al., 1998), as is also the case with species such as *Neisseria gonorrhoeae* (O'Rourke and Spratt, 1994) and *Bacillus subtilis* (Istock et al., 1992), whose population genetic structure has been called panmictic (Maynard Smith et al., 1993). In panmictic species, recombination is sufficiently frequent and unifying selection sufficiently rare that clonal groupings are not expected to survive except over relatively short periods, such as in cases of transmission within families (Suerbaum et al., 1998) or between sexual contacts (O'Rourke et al., 1995). Furthermore, numerous distinct alleles are generated as a result of repeated genetic exchange between divergent lineages. Because most alleles from independent strains of *H. pylori* are unique, it initially seemed unlikely that multilocus sequence typing (Macnab et al., 1998) would reveal any clonal groupings within *H. pylori*. However, the results presented here show that widespread clonal groupings can be discerned even within a largely panmictic species such as *H. pylori*. We note that a widespread clone has also been found within the supposedly panmictic *N. gonorrhoeae* (Gutjahr et al., 1997) and suggest that, with time, more clonal groupings will be

Table 5. Clone 2, a group of three or four strains in which some gene fragments cluster significantly.

<table>
<thead>
<tr>
<th>Gene</th>
<th>ureI</th>
<th>mutY</th>
<th>stdp</th>
<th>cagA(N)</th>
<th>cagA(C)</th>
<th>pps</th>
<th>yphC</th>
<th>vacA</th>
<th>stpA</th>
<th>stpC</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position (kb)</td>
<td>75</td>
<td>154</td>
<td>184</td>
<td>580</td>
<td>581</td>
<td>666</td>
<td>886</td>
<td>939</td>
<td>1197</td>
<td>1355</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strain</th>
<th>H12</th>
<th>5590</th>
<th>CC28</th>
<th>J166</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of strains</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bootstrap (%)</td>
<td>56</td>
<td>90</td>
<td>68</td>
<td>55</td>
</tr>
</tbody>
</table>

Details are as in Table 4.
recognized within other largely panmictic bacterial species. Both the Asian clone and clone2 are currently represented by only a few isolates, and experiments with additional bacteria from additional geographical regions and those that lack the \textit{cag} pathogenicity island are needed to better define their variability and other properties. However, the existence of these weakly clonal groupings seems inescapable, because sequence polymorphisms in multiple housekeeping genes scattered around the chromosome yielded relatively consistent results. These results also demonstrate the power of multilocus sequence analysis for the recognition of clonal groupings, even within panmictic species.

**Clonal groupings versus selection for particular alleles**

The \textit{cagA} gene is in a pathogenicity island that is associated with virulence. Large sequence differences distinguish \textit{cagA} gene fragments from Asian strains versus others (van der Ende \textit{et al.}, 1998) (Fig. 3). The synonymous sequence diversity ($K_s$) of \textit{cagA} is not unusual in comparison with other chromosomal genes within \textit{H. pylori} (Table 3), and the age of all these genes may be similar. Furthermore, the \textit{cagA} sequences from the Asian clone are not particularly uniform (Fig. 3), arguing against a recent selective sweep of the \textit{cag} pathogenicity island, such as has been invoked to explain the relative lack of sequence polymorphism within the \textit{gapA} gene of \textit{Escherichia coli} (Guttman and Dykhuizen, 1994). The alleles of five of the seven housekeeping gene fragments clustered together significantly within at least two of the six Asian strains (Fig. 2, Table 4). It therefore seems likely that the different \textit{cagA} or \textit{vacA} populations in Asia and other continents (Ito \textit{et al.}, 1998; van der Ende \textit{et al.}, 1998) reflect an old clonal expansion. Similar conclusions have been drawn for \textit{Mycobacterium tuberculosis} (Van Soolingen \textit{et al.}, 1995), in which the insertion sites for IS6110 were conserved in a considerable proportion of strains isolated from East Asians. Geographical partitioning has also been documented for type b \textit{Haemophilus influenzae} (Musser \textit{et al.}, 1990), for which
unusual electrophoretic types were commonly isolated in certain countries to which recent human migration has been rare.

We have combined the data on cagAC presented here with that published elsewhere (van der Ende et al., 1998) and with several unpublished cagAC sequences (data not shown). The three distinct branches in the cagA dendrogram (Fig. 3) remain unchanged with these 75 sequences. Analysis of the larger set of sequences has also indicated that distinct branches of the cagA phylogenetic dendrogram can largely be accounted for by clustered amino acid polymorphisms that are uniform within each individual branch (data not shown). Specific amino acid polymorphisms might reflect different, specific interactions between the CagA protein and physiologically or genetically distinct human hosts, and selection for such polymorphisms may have helped to maintain clonal groupings despite frequent recombination. Elucidation of such interactions will depend on identifying the function of the CagA protein in human colonization and disease. The data presented here also indicate that particular allelic variants in the signal peptide of the vacA gene are associated with each of the two clonal groupings. The different regions of vacA seem to correlate with different target cell specificities of the toxin (Pagliaccia et al., 1998).

Age of H. pylori

How old is H. pylori and how old is the Asian clone? The mean $K_s$ values of the various genes in Table 3 and similar values for vacA, flaA and flaB (Suerbaum et al., 1998) are higher than for most housekeeping genes in Drosophila melanogaster, E. coli and N. meningitidis. This observation suggests that H. pylori has a long history, possibly of millions of years. Sequence variation within the Asian clone was less than within the entire species, but sufficiently high that the Asian clone of H. pylori may have accompanied Homo sapiens when they colonized Asia some 40,000 or more years ago. Alternatively, a strain variant of H. pylori that is particularly well adapted to East Asians might have spread through that population more recently, displacing bacteria that were formerly present. Apparently, subsequent recombination has not been sufficiently frequent to obscure the evidence for clonal descent. The age of clone2 may be similar to that of the Asian clone because the sequence variability of the two sets of strains was similar (Fig. 2).

Free recombination versus clonality

The data presented here provide evidence for both extensive recombination and clonality. One possible explanation for such mutually incompatible models is that their relative importance for H. pylori differs with the geographical region. For example, lower $H$ ratios were observed with strains from South Africans (including CC28, which belongs to clone2) than with strains isolated in Canada and Germany (Suerbaum et al., 1998). Not all of H. pylori is necessarily clonal: remnants of clonal descent were not detected among the 10 strains that were not assigned to the Asian clone or clone2. Possibly, the clonal members differ in the frequency with which genes transferred from other strains of H. pylori are fixed. Such differences are known in other species, such as N. meningitidis, in which seven distinct clonal groupings of hypervirulent bacteria exist (Maiden et al., 1998) even though recombination is frequent (Morelli et al., 1997) and many other isolates are panmictic (Maynard Smith et al., 1993). Finally, the spread of H. pylori from one human host to the next may be sufficiently slow that insufficient time has elapsed for recombination to remove all traces of clonal descent.

Frequencies of recombination

We used both the homoplasy test and compatibility matrices to ensure that our conclusions on recombination did not reflect the limitations of a single method. The results with both methods were consistent with two exceptions, namely cagAN, which is clonal according to the $H$ ratio, but seems from compatibility matrices to recombine frequently, and $trpC$, which has undergone less recombination than the other housekeeping genes based on its $H$ ratio, but not according to compatibility matrices. Possibly, these discrepancies can be accounted for by the high $K_s$ values for both exceptional genes. Compatibility matrices evaluate all informative sequence variation and would be affected by the higher level of non-synonymous variation in these gene fragments, whereas the homoplasy test only uses synonymous, informative polymorphisms. Both methods agreed that cagAC was highly clonal and that recombinants were very rare. Thus, different genes (or gene fragments) within one chromosome of a bacterial species may be associated with different degrees of recombination.

In other species, elevated levels of polymorphism have been attributed to linkage to gene clusters that are imported from other species (gnd linked to rib) (Bisero et al., 1991) or to diversifying selection (Smith et al., 1995; Boyd et al., 1997). However, the genes we chose for sequencing encode cytoplasmic enzymes, and they are not linked to genes encoding outer membrane or secreted proteins that might be under selection. We have also argued above that the evidence does not support a selective sweep having caused the clonal descent associated with the cagA gene. Thus, the mechanisms responsible for different degrees of recombination associated with individual genes in H. pylori remain to be elucidated.
Summary

Two clonal groupings, the Asian clone and clone2, were detected in a collection of H. pylori from several parts of the world. These clonal groupings are widespread and have probably existed for a long time. Recombination seems to have occurred on a global scale for most genes studied, but has not totally disrupted the relationships with the clonal groupings. The results presented here should stimulate further extensive analyses of the geographical specialization and evolution of this species.

Experimental procedures

Bacterial strains

Nineteen strains were chosen from diverse ethnic groups and countries to represent the geographical diversity of H. pylori (Table 1). The collection deliberately included many strains that had already been used in previous analyses. Because of our intention of examining sequence variation of the cagA gene, we only included strains thought to be CagA+. The 19 strains were purified by single-colony isolation before further analysis. These strains, or their DNAs, are available from the individual sources shown in Table 1.

DNA, genes, PCR products and sequences

DNA was purified from cultures of H. pylori by the CTAB method (Ausubel et al., 1994) or by other standard methods appropriate for PCR amplification of DNA fragments and sent in dried form to each of the laboratories of the consortium. We chose seven housekeeping genes for sequencing (Table 2), using the criteria that these genes be both widely separated on the chromosome of the strain (26695) whose genome has been sequenced (Tomb et al., 1997) and not adjacent to genes encoding putative outer membrane, secreted or hypothetical proteins that might be under selective pressure. Based on the sequences in the genome of strain 26695, amplification and

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name (orientation)</th>
<th>Purpose</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpA</td>
<td>AtpA1(+) AtpA2(-)</td>
<td>A, S</td>
<td>GCTTAAATGGTGTGATGTCG</td>
</tr>
<tr>
<td></td>
<td>AtpA3(+) AtpA4(-)</td>
<td>S</td>
<td>AAGTTCGCGAGCGCTATAGC</td>
</tr>
<tr>
<td>cagA</td>
<td>cagA1(+)</td>
<td>A, S</td>
<td>AGGATTAGACGCTATACAC</td>
</tr>
<tr>
<td></td>
<td>cagA2(-)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
<tr>
<td></td>
<td>cagA3(+)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
<tr>
<td></td>
<td>cagA4(-)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
<tr>
<td></td>
<td>cagA5(+)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
<tr>
<td></td>
<td>cagA6(-)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
<tr>
<td></td>
<td>cagA7(+)</td>
<td>A, S</td>
<td>GCCACATGCTGACGTGAGCC</td>
</tr>
</tbody>
</table>

Table 6. Oligonucleotide primers.
sequencing primers were designed such that complete double-stranded sequences could be obtained with a single sequencing run in each direction. For several gene fragments, sequence diversity prevented amplification from all strains with a single pair of primers. Additional primers were designed from the available sequences and, where necessary, after cloning and sequencing PCR amplicons that had been amplified inefficiently. The complete set of primers used for the seven housekeeping genes is shown in Table 6. Two gene fragments of cagA (cagAN, which corresponds to the N-terminus, and cagAC, which is slightly C-terminal of cagAN) and a vacA gene fragment were also sequenced to allow comparison with previous results (van der Ende et al., 1998; van Doorn et al., 1999). The primers used to amplify these genes are also included in Table 6 for convenience, although they have been published previously (Pan et al., 1997; Suerbaum et al., 1998; van der Ende et al., 1998; van Doorn et al., 1999). Automated sequencing of PCR products from both strands was performed using standard protocols, and sequences were deposited in GenBank under the designations listed in Table 2. vacA's and m typing was performed by hybridization as described previously (van Doorn et al., 1998b).

Phylogenetic and other analyses

Sequence alignments in GCG MSF format were converted to MEGA format using the program EMBL and evaluated with the homology test (Maynard Smith and Smith, 1990), using the HOMOLOGY program (Suerbaum et al., 1998). Phire and HOMOLOGY are both available by anonymous ftp at http://novell-del-valle.rz-berlin.mpg.de/software/. Homology rates are the means of five repetitions of the homology test. Mean Ks and Kn distances, using Jukes-Cantor corrections (Jukes and Cantor, 1969), were calculated using DNAP 2.2 (Rozas and Rozas, 1997). Compatibility matrices and mean percentage compatibility values were calculated using the program RETICULATE (Jakobsen and Easteal, 1996). Pairwise distance matrices based on numbers of different nucleotides (p-distance) were calculated using the MEGA program (Kumar et al., 1993) and normalized to a maximum of 1.0 for each matrix. These matrices were averaged (Excel, Microsoft) and used to generate an UPGMA tree (Statistica, Statsoft). Bootstrap tests (Felsenstein, 1985) (500 repetitions of neighbour-joining trees based on distances with Jukes-Cantor corrections were performed using MEGA. Maximum-likelihood tree and graphical outputs were calculated using the program PHYLIP (http://www.mikro.biologie.tu-muenchen.de). Percentage GC was calculated using the program concov (http://molbiol.ox.ac.uk/cu/).

Acknowledgements

We gratefully acknowledge the expert assistance of Monique Feller, Susanne Friedrich, Ricardo Sanna, Kerstin Zurch and grant support from the Deutsche Forschungsgemeinschaft (Su 133/3-1, Ac 56/10-1) and the US Public Health Service (NlH grants AI36166, DK43023 and HG00820). We thank the following for H. pylori strains studied in this work: A. J. Parkinson (96-228, 96-252), P. H. Gilman (HP1), A. Lee (SS1), C. Garcia and O. Torres (#12) and E. Kunstmann (BO255, CC28).

References

Bamford, K.B., Bickley, J., Collins, J.S., Johnston, B.T., Potts, S., Boston, V., et al. (1993) Nucleotide sequences and mean percentage compatibility values were calculated using the program RETICULATE (Jakobsen and Easteal, 1996). Pairwise distance matrices based on numbers of different nucleotides (p-distance) were calculated using the MEGA program (Kumar et al., 1993) and normalized to a maximum of 1.0 for each matrix. These matrices were averaged (Excel, Microsoft) and used to generate an UPGMA tree (Statistica, Statsoft). Bootstrap tests (Felsenstein, 1985) (500 repetitions of neighbour-joining trees based on distances with Jukes-Cantor corrections were performed using MEGA. Maximum-likelihood tree and graphical outputs were calculated using the program PHYLIP (http://www.mikro.biologie.tu-muenchen.de). Percentage GC was calculated using the program concov (http://molbiol.ox.ac.uk/cu/).


