Analysis of geospecific markers for *Helicobacter pylori* variants in patients from Japan and Nigeria by triple-locus nucleotide sequence typing

Robert J. Owen,¹ Jacqueline Xerry,¹ Takuji Gotada,² Greg Naylor³ and David Tompkins⁴

¹Laboratory of Enteric Pathogens, Specialist and Reference Microbiology Division, Health Protection Agency, London NW9 5HT, UK
²Department of Endoscopy and Gastrointestinal Endoscopy, National Cancer Centre Hospital, Tokyo 104-0045, Japan
³The Centre for Digestive Diseases, The General Infirmary at Leeds, Leeds LS1 3EX, UK
⁴Health Protection Agency Laboratory, York Road, Leeds LS15 7TR, UK

Human migrations and geographical separation over long periods may have resulted in ecologically distinct populations of *Helicobacter pylori* infecting individuals in different continents. This study used nucleotide sequence analysis with the aim of defining population-specific genomic motifs in isolates from East Asian and African dyspeptic patients. Sequences of internal fragments (542–627 bp) of three housekeeping genes (*ureI*, *ahpC* and *atpA*) were analysed for 85 isolates from individuals in Japan and China (30 isolates), Nigeria and South Africa (14 isolates), the United Kingdom (32 isolates), and nine miscellaneous reference strains. Phylogenetic analyses showed a high degree of intra-set relatedness amongst sequences from the Japanese and Nigerian isolates, with each robustly segregated as distinct lineages irrespective of *cagA* presence and *vacA* allelic type. All strains had unique combined sequence types except for identical paired (antrum/corpus) isolates. Population-specific polymorphisms were identified within each gene which were combined to provide unique motifs defining the Japanese and Nigerian regional populations. The alleles were present at variable frequencies in UK and South African isolates. The findings provide unique evidence of positive selection for conserved nucleotide sites linked to the geographical separation in Japan of a strain subpopulation for which we propose the designation *H. pylori* geovar ‘orientalis’.

INTRODUCTION

*Helicobacter pylori*, a Gram-negative microaerophilic bacterium, is one of the most prevalent bacterial infectious agents amongst humans worldwide, with rates exceeding 40 % in many countries (Pounder & Ng, 1995). Infections are chronic and, after acquisition in early childhood, are likely to be lifelong unless eradicated with antibiotics. Most infected individuals remain asymptomatic but in some the micro-organism is an important causal factor in the development of peptic ulcer disease and gastric low-grade mucosal-associated lymphoid tissue lymphoma as well as a key risk factor in the development of gastric cancer (Dunn *et al.*, 1997; Calam & Baron, 2001). Specific disease-associated strain markers have not yet been defined, and the factors determining onset of chronic disease remain unclear but could be a complex combination of host, environmental and bacterial interactions.

Isolates of *H. pylori*, irrespective of geographical location, have identical phenotypes as defined by conventional bacteriological tests (Owen *et al.*, 2001b) although there is evidence that strains can differ in some expressed features such as vacuolating cytotoxin activity (Forsyth *et al.*, 1998), lectin reaction patterns (Hynes *et al.*, 2002) and interleukin-8 induction in gastric epithelial cells (Owen *et al.*, 2003). In contrast, strains of *H. pylori* from unrelated individuals exhibit an unusually high level of genomic diversity as indicated by various fingerprinting techniques (Owen *et al.*, 2001b). Multilocus sequence analyses indicate that the species has a panmictic (non-clonal) population structure due to frequent horizontal transfer and free recombination of metabolic enzyme genes, and of pathogenicity-associated genes such as *cagA* and *vacA* (Suerbaum *et al.*, 1998; Achtman *et al.*, 1999). These genetic events are thought to
occur in individual patients during mixed gastric colonization by unrelated strains over several years (Falush et al., 2001). Nevertheless, analysis of sequence variability also provides evidence of two weakly clonal groupings (Asian, and African/Afro-American group) in strains from different parts of the world despite extensive inter-strain recombination masking evidence of phylogenetic relatedness (Achtman et al., 1999). Sequence analysis of housekeeping and pathogenicity-associated genes, mainly on isolates from Switzerland but with representatives from China and South Africa, also demonstrated that the species existed as a recombinant population, but no evolutionary relationships between strains could be discerned in that strain set (Solca et al., 2001). More recently, evidence that H. pylori can be divided into seven populations and subpopulations with distinct geographical distributions was established from a more extensive analysis of multilocus data implementing a Bayesian approach for deducing population structure (Falush et al., 2003). The possibility that geographical separation of H. pylori may have resulted in distinct populations evolving on different continents is also suggested by analyses of the virulence-associated cagA and vacA genes, where differences were reported between isolates originating from China and Japan compared to those from the United States and Europe (van der Ende et al., 1998; Ito et al., 1997; Pan et al., 1998; van Doorn et al., 1998; Yamaoka et al., 1998). Geographically associated characters have also been demonstrated for urease subunit genes (Campbell et al., 1997), and for an outer-membrane protein gene (HP0638) (Ando et al., 2002).

H. pylori from different parts of the world, nevertheless, remain difficult to stratify in relation to disease severity. Our study focuses on patient isolates from Japan to define East Asia population-specific markers by comparison with isolates from sub-Saharan Africa, represented by Nigeria and South Africa, and from the United Kingdom, represented by England and Northern Ireland. Sequence diversity was examined in core fragments of two established housekeeping genes (ureI and atpA), and at a previously uninvestigated locus (ahpC), in H. pylori from unrelated individuals who were selected wherever possible as members of the local indigenous population. We aimed also to determine the frequencies of the specific East Asian markers in the other strain subpopulations.

METHODS

Bacterial strains. We studied 85 isolates of H. pylori from 70 patients in three widely separated geographical regions of the world. The isolates were all from gastric biopsies of unrelated patients undergoing routine upper gastrointestinal endoscopic investigation for various presentations between 1990 and 2002. The East Asian set comprised 27 Japanese (Tokyo) isolates collected as part of the UK/Japan Gastritis Study from 16 patients (eight males/eight females) aged between 23 and 79 years (mean 55-6 years). For eleven patients, paired (antrum and body isolates) were tested. In addition, one isolate from a dyspeptic patient in China (Jinan City, Shandon Province), and isolates from two Chinese patients who were UK residents, were included. The African set comprised eight isolates from patients in Nigeria (Maiduguri, Borno State) provided by Dr C. Holcombe (Holcombe et al., 1992, 1994), and a further six strains that originated from patients (Cape-coloured population) in South Africa (Cape Town) provided by Dr A. Lastovica (Owen et al., 1992). The UK set comprised 23 isolates from 18 patients (white Caucasian and UK born) with paired (antrum and corpus) isolate sets from five individuals, and nine isolates from resident unrelated UK-Asians of Indian, Pakistani and Bangladeshi origin who were born outside the UK.

Nine geographically diverse but internationally recognized reference strains of H. pylori were also examined. Seven cultures were obtained (in lyophilized form) from the National Collection of Type Cultures (NCTC): NCTC 11637 (type strain, Australia); NCTC 11638 (Australia); NCTC 13081 (strain TX30a, USA); NCTC 13206 (CCUG 38770, antibiotic susceptibility control strain, Belgium) (Glupczynski et al., 2001); NCTC 13207 (CCUG 38772, antibiotic susceptibility control strain, France) (Glupczynski et al., 2001); NCTC 12455 (strain 26695, genome-sequenced strain, England) (Tomb et al., 1997); and NCTC 12491 (formerly type strain of Helicobacter nemestriniae, USA) (Suerbaum et al., 2002). Cultures of the genome-sequenced strain J99 (USA) (Alm et al., 1999) and of strain SSI (Australia) (Lee et al., 1997) were kindly provided by Professor D. Taylor (University of Alberta, Canada) and Dr S. Rypkema (National Institute for Biological Standards and Control, Potters Bar, UK), respectively.

Culture conditions. All strains of H. pylori from biopsy specimens and allied species were cultured for 2–3 days on Columbia agar base (Oxoid) containing 10% (v/v) defibrinated horse blood, and incubated at 37°C under microaerobic conditions (4% oxygen, 5% hydrogen, 5% carbon dioxide and 86% nitrogen) in a variable atmosphere incubator (Don Whitley Scientific). Identity of H. pylori was confirmed by Gram staining, and catalase, cytochrome oxidase and urease tests. Stock cultures were preserved on glass beads at −80°C (Microbank system; Pro-Lab Diagnostics) or over liquid nitrogen.

DNA preparation, PCR and sequencing. Genomic DNA was amplified by the method of Wilson (1987). Diluted DNA (100 ng) was used to amplify an internal fragment from three genes (ureI, atpA and ahpC) which were selected because of their potential roles in core cytoplasmic metabolic housekeeping activities and not in encoding outer-membrane proteins or secreted proteins that might be under greater selection pressure in a stressful environment such as the gut mucosal surface. The ureI locus, which was included previously in a panel of seven housekeeping genes used for a population genetic study (Achtman et al., 1999), is part of the H. pylori urease gene cluster and encodes a urea accessory (transporter) protein thought to be an integral cytoplasmic membrane protein forming a proton-gated urea-channel-regulating cytoplasmic urease (Weeks et al., 2000; Mobley, 2001). The following primers were designed for amplification and sequencing of a 585 bp fragment within the ureI sequence (HP0071) of H. pylori strain 26695: forward primer (ureI5′-GGAGGAAAAGGCAATGCC-3′) and reverse primer (ureI3′-CTAAACGCTCTATGATCA-3). These primers were used as alternatives to those described by Achtman et al. (1999), which were initially tested but did not amplify ureI from all our isolates. The target region for these primers corresponded closely to those described by Achtman et al. (1999).

The atpA gene encodes an ATP synthase F1 alpha homologue forming the catalytic portion of the multisubunit enzyme found in the cytoplasmic membrane. The enzyme has a key function in the synthesis of ATP and may have a role in adaptation by H. pylori to acid stress by proton translocation (McGowan et al., 1997). Primers for amplification and sequencing of a 627 bp region within the atpA (HP1134) sequence were as described by Achtman et al. (1999).
The ahpC gene encodes the AhpC subunit of an alkylhydroperoxide reductase – a cysteine-based peroxidase homologue involved in the defence against oxidative stress. It is present in *H. pylori* strain 26695 (HP1563) as well as in at least eight other species of *Helicobacter* (Lundstrom et al., 2001), and although not used in previous sequence typing studies, it was evaluated in the present study as a potential novel locus to provide population-specific markers. The forward primer (HP26F) 5′-TTAGTACAAACTGTGCCC-3′, and reverse primer (HP26R) 5′-GCTTTACCTCCCTATCCGC-3′) were used to amplify and to sequence a 542 bp fragment as described by Lundstrom et al. (2001).

**PCR and nucleotide sequencing.** The three sets of primers were synthesized commercially (MWG BIOTECH) and PCR amplification reactions were performed in a total volume of 50 μl containing 100 ng of diluted template DNA, 1·5 mM MgCl₂, 0·05 mM each deoxynucleotide (dATP, dCTP, dGTP and dTTP), 0·4 mM each oligonucleotide primer, 0·2 μl (1 U) Taq polymerase (invitrogen) and 5 μl of 10× buffer provided by the manufacturer. Automated sequencing from both strands of PCR products of the internal fragments was then performed by standard protocols either in house using a CQ2000 sequencer (Coulter Beckman) or commercially (MWG BIOTECH). Use of these six primer sets enabled complete double-stranded sequences to be obtained from a single sequencing run in each direction. Twenty-one sequences representing the range of diversity of loci within the East Asian (four strains) and Nigerian (two strains) population sets have been deposited in EMBL with accession numbers AJ583137–AJ583157.

**Phylogenetic and polymorphism analyses.** The *urel*, *ahpC* and *atpA* sequences of 76 isolates and seven reference strains of *H. pylori* as well as additional *H. pylori* sequences deposited in GenBank for the two genome-sequenced strains (26695 and J99) and nine other miscellaneous strains (Japan, 2 strains; China, 2 strains; Thailand, 2 strains; Gambia, 2 strains; and South Africa, 1 strain) were aligned using CLUSTAL W as implemented in BioEdit (v 5.0.9) (Hall, 1999). Sequences of the three loci for *Helicobacter mustelae* NCTC 12198 were included in the analyses to provide an outlier to root trees. The sets of aligned sequences for each gene were analysed by the neighbour-joining method of Saitou & Nei (1987) using the nucleotide substitution model of Dukes & Cantor (1969), and trees were produced in TREECON (van de Peer & De Wachter, 1994). Bootstrap sampling analyses were performed with 100 resampled datasets to define confidence limits in the estimated phylogenies. Phylogenetic analyses were also performed using a dataset in which the three sequences (1754 bp) were concatenated in the order *urel*/*atpA*/*ahpC* for each strain. Further analysis of nucleotide sequences was performed with DnaSP v 3.51 (Rozas & Rozas, 1995) to determine base composition, the number of polymorphic sites including the number of parsimony-informative sites (PI), and deduced amino acid sequences. Also calculated were the proportions of synonymous substitutions (K₀), defined as the percentage of mean differences between pairs of strains at synonymous sites (mutations not resulting in amino acid substitutions) and the proportion of non-synonymous substitutions (K₁) defined as the percentage of mean differences between pairs of strains at non-synonymous sites (mutations resulting in amino acid substitutions). Low K₀/K₁ ratios (<1·0) indicated frequent non-synonymous site variation. Sequences were screened to identify single and multiple nucleotide polymorphisms (NPs) with population-specific associations.

**Assessment of *H. pylori cagA* and *vacA* genotype status.** The primers and PCR conditions for the two assays for cagA (a marker for the 3′ end of the cag pathogenicity island and for the cagl region) were as described previously using the F1/B1 primers and the D008/R008 primer sets (Slater et al., 1999; Owen et al., 2001a). Vacuolating cytotoxin (*vacA*) genotyping based on signal and mid-region alleles was performed as previously described (Atherton et al., 1999; Owen et al., 2002).

**RESULTS**

**Phylogenetic analysis of *H. pylori* diversity – general features**

Internal fragments of *urel*, *atpA* and *ahpC* were amplified and sequenced for 83 cultures of *H. pylori*. None of the sequences contained gaps or insertions. The phylogenetic trees derived from analysis of the three concatenated gene sequences of those strains and the two genome-sequenced strains (26695 and J99) obtained from GenBank showed that all were closely related, with a high level of similarity of ≥95% over 1754 nt. A distinct feature of rooted and unrooted trees was the segregation of strains into distinct lineages corresponding to geographical location. In the tree rooted to *H. nesemetrae* (Fig. 1), 30 isolates from 19 unrelated individuals forming a distinctive East Asian branch (clade A) containing isolates from patients in Japan (n = 27 isolates); and China-mainland (n = 3 isolates) with intragroup similarities of 96·9–99·9%. Members of nine of the 11 paired (antrum/body) isolate sets had identical sequences (not shown on the tree). One patient had two closely related strains (H2695 and H2737) whereas another patient had unrelated strains (H2735 and H2778). All patients were infected by strains with different sequences.

The eight Nigerian isolates also formed a discrete geographical cluster (clade B), with intragroup similarities of 98·0–99·1%, that was located separately from the East Asian isolates within the diversity continuum. Inter-clade similarities were 96·2–96·9%. Nearest neighbours were NCTC 11638 and Tx30a. No definable phylogenetic substructure was evident amongst the geographically heterogeneous strains located between clades A and B in the tree. Interestingly, the genome-sequenced strain J99, isolated in the United States, was more closely linked to four South African isolates.

The three phylogenetic trees obtained from separate analyses of the *urel*, *atpA* and *ahpC* loci confirmed the distinctiveness of the East Asian and the Nigerian isolates. By contrast, the detailed structures and internal order of strains within the resultant trees, especially for those of the UK and South Africa isolates, were not completely concordant with each other or with the composite tree (results not presented). For the *urel* and *atpA* analyses, additional sequences from GenBank showed that one Nigerian isolate (H2495) from clade B was closely linked to strain 5596/Gambia, which was a representative of clone 2 (Achtman et al., 1999). Furthermore, three strains (R29/Japan, 97–42/China and 88–8/Thailand), part of the ‘Asian’ clone (Achtman et al., 1999), were closely related to clade A. Two other East Asian isolates (F32/Japan and 88–39/Thailand) fell outside clade A and were excluded from...
further analyses as ethnic details of patients could not be confirmed.

**Identification of *H. pylori* nucleotide and amino acid sequence types**

The numbers of *H. pylori* representing each geographical group and results on nucleotide sequence type (NST) and amino acid sequence type (AST) derived for the three genes (*ureI*, *atpA* and *ahpC*) are listed in Table 1. There was a high degree of sequence diversity amongst alleles from all isolates. Every patient was infected by a strain with a unique NST except for the paired antrum/corpus isolates from eleven Japanese and from five UK patients, where paired strains had an identical ST. In the UK set, there were 17 NSTs for both *ureI* and *atpA* compared to 18 for *ahpC*. The

---

**Fig. 1.** Phylogenetic tree constructed from analysis of nucleotide sequences of *H. pylori* reference strains and isolates. The analysis was based on a composite sequence (1754 bp) for each strain by combination of data on the three loci in the order *ureI*, *atpA* and *ahpC*. The isolate laboratory number is indicated on the tree followed by the country of origin of the infected individual. Sequences of strains representing clades A and B have been deposited in the EMBL database (see Methods), and sequences of other strains in public databases are listed by Owen & Xerry (2003). Bootstrap values of 60% and higher (expressed as percentages of 100 replications) are shown at key branching points. The tree was rooted using *H. nemestrinae* NCTC 12491, a strain of *H. pylori* originating from an animal (primate) host.
Table 1. *H. pylori* isolates and sequences analysed according to geographical location, with nucleotide sequence type (NST) frequencies and corresponding amino acid sequence type (AST) in parentheses, and pathogenicity-associated markers

<table>
<thead>
<tr>
<th></th>
<th>East Asian set</th>
<th>African set</th>
<th>UK set</th>
<th>Reference set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Japan*</td>
<td>China/Thailand†</td>
<td>Nigeria</td>
<td>S. Africa‡</td>
</tr>
<tr>
<td>Isolates in each set</td>
<td>29</td>
<td>5</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>No. of NST (AST) types</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ureI</td>
<td>18 (9)</td>
<td>5 (4)</td>
<td>8 (5)</td>
<td>6 (5)</td>
</tr>
<tr>
<td>atpA</td>
<td>18 (5)</td>
<td>5 (4)</td>
<td>8 (3)</td>
<td>6 (3)</td>
</tr>
<tr>
<td>ahpC</td>
<td>18 (6)</td>
<td>3 (3)</td>
<td>8 (2)</td>
<td>6 (3)</td>
</tr>
<tr>
<td>No. of isolates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cagA present</td>
<td>28‖</td>
<td>5</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>vacA s1m1</td>
<td>29</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>vacA s1m2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>vacA s2m2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*Paired isolates were tested from eleven patients and in nine of these pair members were identical.
†GenBank data included for two Japanese and two Thai strains.
‡Isolates were all from the Cape-coloured population.
§Paired isolates were tested from five patients and members of all pairs were identical.
‖One patient had cagA+ and cagA− variants at different sites.
†This culture was a mixture of m1 and m2 genotypes.

The ureI sequences were screened for population-specific nucleotide polymorphisms and the results, summarized in Table 3, show seven diallelic substitutions that were characteristic of the core East Asian set (clade A) and of the Nigerian set (clade B), with individual population frequencies of >96%. At three sites the substitutions were all non-synonymous. Comparison of the UK-white and UK-Asian strains did not indicate polymorphisms that were specific for either subpopulation. However, the UK isolates had unique ureI haplotypes (combinations of linked polymorphisms) that were distinct from those of the East Asian and Nigerian isolates, with four (160/C, 221/G, 237/G and 252/T) of the seven individual polymorphisms matching those of the corresponding Nigerian allele. Likewise, the South African isolates had unique combinations of substitutions compared to the Nigerian isolates indicating some regional variation in Africa, although the Gambian strain substitutions matched those from Nigeria.

**Definition of population-specific nucleotide polymorphisms**

**Nucleotide polymorphisms in ureI**. The relevant sequence-derived parameters for each of the gene fragments, such as base composition, number of polymorphic sites, and mean K/S/KA values, are listed in Table 2. The ureI fragment G+C content of 43.7 mol% was conserved across the geographical groups whereas the frequencies of polymorphic sites varied over a narrow range from 4% (East Asian-others) to 9% (UK-white) with a mean of 6%. Substitutions were randomly distributed over the fragment sequences. The mean K/S/KA ratios were also conserved and varied from 11·2 (UK-white) to 17·4 (Nigerian), with a pooled mean of 14·2.

**Nucleotide polymorphisms in ahpC**. The G+C contents of the ahpC sequences showed distinct differences between geographical groups, with a value of 41·3±0·2 mol% for the East Asian set whereas the African set was lower at 40·6±0·1 mol% and the UK set was markedly higher at 46·3±0·1 mol%. Polymorphism frequencies ranged from 3% for the Nigerian strains to 10% for the UK-white strain subset, with a mean frequency of 6%. The mean K/S/KA ratios varied markedly between sets; most striking was the sixfold difference between the Nigerian sequences at 90·1 compared to the Japanese at 14·6. The UK-white sequences had the second highest ratio at 65 whereas the South African and UK-Asian sequences had intermediate ratios of 45 and 38 respectively. The...
histidine to asparagine; at site 197, valine to alanine; and at site 221, serine to asparagine. The non-synonymous mutations indicated resulted in the following amino acid substitutions: at site 160, histidine to asparagine; at site 197, valine to alanine; and at site 221, serine to asparagine.

*N. pylori* ureI

Analysis of frequencies of population-specific nucleotide polymorphism (PSNP) in the *H. pylori* ureI locus (HP0071) of isolates (including multiples from individuals) from different countries

<table>
<thead>
<tr>
<th>Features of each locus</th>
<th>East Asian†</th>
<th>African</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Japanese</td>
<td>Others</td>
<td>Nigerian</td>
</tr>
<tr>
<td><em>ureI</em> (585 sites)</td>
<td>(n = 28)</td>
<td>(n = 4)</td>
<td>(n = 8)</td>
</tr>
<tr>
<td>G + C content (mol%)</td>
<td>43·7</td>
<td>43·7</td>
<td>43·8</td>
</tr>
<tr>
<td>No. of polymorphic sites</td>
<td>39 (7%)</td>
<td>21 (4%)</td>
<td>28 (5%)</td>
</tr>
<tr>
<td>No. of PI sites*</td>
<td>21</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Mean <em>Ks/Ka</em></td>
<td>14·46</td>
<td>12·25</td>
<td>17·35</td>
</tr>
</tbody>
</table>

| *ahpC* (542 sites)     | (n = 28)    | (n = 1) | (n = 8)  | (n = 6)  | (n = 19) | (n = 9)  |
| G + C content (mol%)   | 41·5        | 41·1    | 40·5    | 40·7    | 46·4     | 46·2     |
| No. of polymorphic sites | 27 (5%)   | NR      | 16 (3%) | 31 (6%) | 52 (10%) | 40 (7%)  |
| No. of PI sites         | 14          | NR      | 7       | 8       | 34       | 21       |
| Mean *Ks/Ka*           | 14·6        | 16·0    | 90·59   | 45·26   | 65·44    | 38·11    |

| *atpA* (627 sites)     | (n = 28)    | (n = 4) | (n = 8)  | (n = 6)  | (n = 19) | (n = 9)  |
| G + C content (mol%)   | 46·3        | 46·2    | 46·1    | 46·1    | 46·4     | 46·2     |
| No. of polymorphic sites | 43 (7%)   | 32 (5%) | 32 (5%) | 32 (5%) | 52 (8%)  | 40 (6%)  |
| No. of PI sites         | 20          | 4       | 16      | 19      | 34       | 21       |
| Mean *Ks/Ka*           | 56·6        | 24·4    | 44·8    | 41·16   | 65·44    | 38·11    |

*PI, parsimony informative.
†Others comprises sequences for China (three) and Thailand (one). Analysis excluded GenBank sequences for strains F32/Japan and 88–39/Thailand that were found to be atypical in a pilot phylogenetic analysis and ethnic origin details could not be confirmed.

High ratios indicated a greater degree of conservation of AhpC amino acid sequences in the Nigerian and UK-white populations. The analysis identified 14 di-allelic substitutions in *ahpC* that were specific for the East Asian and the Nigerian isolate sets (Table 4). For three sites, the substitutions were non-synonymous. None of the UK isolates (white and Asian sets) or the South African isolates had haplotypes exactly matching those of the East Asian or Nigerian strains but overall the latter geographical-specific substitutions were more frequent.

**Nucleotide polymorphisms in atpA.** The G + C contents of the *atpA* sequences were conserved across the different geographical groups (mean 46·2 %) as were the frequencies of polymorphic sites, which showed minimal variation of about 6 %. Mean *Ks/Ka* ratios were diverse and in the order East-Asia (24) < UK-Asian (38) < South African (41) < Nigeria (45) < Japan (57) < UK-white (65). Comparative analysis of *atpA* sequences of the East Asian and Nigerian isolates identified population-specific di- and tri-allelic substitutions present at seven sites and

**Table 2.** DNA polymorphism analysis of *H. pylori* genes according to geographical group (n = no. of sequences analysed)

**Table 3.** Analysis of frequencies of population-specific nucleotide polymorphism (PSNP) in the *H. pylori* ureI locus (HP0071) of isolates (including multiples from individuals) from different countries

<table>
<thead>
<tr>
<th>PSNP identity (7 sites)</th>
<th>Predominant alleles (frequency of 1)</th>
<th>East Asian (n = 32)</th>
<th>Nigerian (n = 8)</th>
<th>Number (frequency) of isolates with East Asian alleles in isolates from:</th>
<th>UK (n = 19)</th>
<th>Asian (n = 9)</th>
<th>All (n = 28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>114</td>
<td>G</td>
<td>A</td>
<td>10 (0·53)</td>
<td>4 (0·44)</td>
<td>14 (0·5)</td>
<td>5 (0·71)</td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>G</td>
<td>A</td>
<td>14 (0·74)</td>
<td>6 (0·67)</td>
<td>20 (0·71)</td>
<td>2 (0·29)</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>A</td>
<td>C</td>
<td>4 (0·21)</td>
<td>0</td>
<td>4 (0·14)</td>
<td>4 (0·57)</td>
<td></td>
</tr>
<tr>
<td>197</td>
<td>A</td>
<td>G</td>
<td>17 (0·89)</td>
<td>7 (0·78)</td>
<td>24 (0·86)</td>
<td>6 (0·86)</td>
<td></td>
</tr>
<tr>
<td>221</td>
<td>A</td>
<td>G</td>
<td>4 (0·21)</td>
<td>0</td>
<td>4 (0·14)</td>
<td>3 (0·43)</td>
<td></td>
</tr>
<tr>
<td>237</td>
<td>A</td>
<td>T</td>
<td>0</td>
<td>1 (0·11)</td>
<td>1 (0·04)</td>
<td>2 (0·29)</td>
<td></td>
</tr>
<tr>
<td>252</td>
<td>C</td>
<td>T</td>
<td>1 (0·05)</td>
<td>0</td>
<td>1 (0·04)</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*The non-synonymous mutations indicated resulted in the following amino acid substitutions: at site 160, histidine to asparagine; at site 197, valine to alanine; and at site 221, serine to asparagine.
one site respectively (Table 5). At three sites, the substitutions were non-synonymous. None of the UK isolates or South African isolates had exact haplotype matches with either the Japanese or Nigerian isolates although the Nigerian alleles were the most frequent. The Gambian strain haplotype matched those from Nigeria.

### Analysis of Japanese polymorphisms in different populations

In total, 28 population-specific polymorphisms for three loci were defined by comparison of the Japanese vs Nigerian populations sets. The data were pooled according to

<table>
<thead>
<tr>
<th>PSNP identity (7 sites)*</th>
<th>Predominant alleles (frequency of 1)</th>
<th>Number (frequency) of isolates with East Asian alleles in isolates from:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>East Asian (n=32)</td>
<td>Nigerian (n=8)</td>
</tr>
<tr>
<td>54</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>69</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>75</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>279</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>447</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>561</td>
<td>A/G</td>
<td>C</td>
</tr>
<tr>
<td>615</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>

*Base substitutions at these positions were all synonymous.
†These were all G substitutions.
frequency of occurrence of the Japanese polymorphisms in the three population groups and the results scored as common (>67%), intermediate (<67% but >33%) and rare (<33%). Typically, the Japanese variants were rare, in particular those in ahpC, in both UK-Asian and South African strains, whereas in the UK-white population the frequencies were more variable but mainly in the intermediate/rare categories.

The cagA and vacA genotype status of geographical strain sets

Overall, the cagA gene was present in 86% of all the strains of H. pylori and for individual population sets the highest frequency of 100% was in the Nigerian isolates (8/8) (Table 1). The commonest overall vacA genotype was s1m1 (67%) and that type was a feature of all Japanese isolates, although the Chinese and Thai isolates were typically s1m2 (4/5) (Table 1). The m2 subtype was more common in the UK isolates (14/24, 58%) compared to the other sets (East Asian, 4/22 patients; Nigeria, 1/8 patients; and South Africa, 2/7 patients). The s2m2 type was less common (11%) – present in the UK set (six patients), in the South African set (one patient), and in the reference set (three patients; SS1, Tx30a and CCUG 38772), whereas the combination s2m1 was not found in any of the strains.

DISCUSSION

The present study aimed to identify specific polymorphic nucleotide markers within the overall genetic variability that might be used to characterize Japanese isolates of H. pylori as a distinct biogeographical subpopulation. Investigation of isolates from different human populations is important in providing insights into the interaction of host and environmental factors and of strain virulence, particularly in countries where there are established differences in the rates of certain gastric diseases, such as gastric cancer. Nucleotide sequence analysis of three housekeeping loci (ureI, ahpC and atpA) of H. pylori showed that identical sequences were rare with the exception of the paired antrum/body isolates from the gastric mucosa of the same individuals, and that almost all other strains had unique alleles at each locus. These findings are consistent with previous investigations into H. pylori diversity from different individuals based on analyses of the same or different loci (Achtman et al., 1999; Solca et al., 2001; Owen et al., 2001a). A striking finding in our study was that the three selected gene sequences of H. pylori from individuals in Japan, a country with high rates of gastric cancer (Lambert et al., 2002), contained specific blocks of polymorphic sites that segregated such isolates and other East Asian isolates (clade A) from isolates in the Nigerian (African) population (clade B), in which gastric cancer rates are lower despite a high frequency of H. pylori infection (Holcombe et al., 1992, 1994; Campbell et al., 2001). Our findings agree with the analysis by Falush et al. (2003) of 370 strains from 27 geographical, ethnic and/or linguistic human groupings that assigned almost all strains from various countries in East Asia to the hspEAsia subpopulation – one of three subpopulations in the modern population designated hpEastAsia. Our East Asian isolates were all independent of those investigated by Falush et al. (2003).

Additional evidence in support of distinctions between the East Asian isolates and those from other geographical regions was apparent from our analysis of other parameters of genetic variability, in particular the proportion of synonymous to non-synonymous substitutions in the target sequences. Synonymous substitutions, particularly at the third codon position, are generally more frequent and so contribute most to divergence between strains. However, they are biologically neutral whereas an increased proportion of non-synonymous mutations indicates increased organism–host interaction and the selection of protein variants – for instance a low Ka/Ks ratio of about 4 for cagA could be interpreted as an indicator of antigenic diversity (van der Ende et al., 1998). In H. pylori, synonymous mutations are estimated to be four times more frequent than non-synonymous mutations (Falush et al., 2001), so marked deviations in frequencies may provide valuable population markers. Interestingly, we found several such differences in the Ka/Ks ratios when a mean value of 24 for bacterial genes was used as a comparative reference (van der Ende et al., 1998). The Japanese isolates had similar ratios (about 14-5) for both ureI and ahpC whereas the ratio for atpA of 56-6 was higher (fourfold) and indicated positive selection pressure to conserve the catalytic site of the AtpA protein in that population. Most striking was the sixfold difference between the ahpC mean Ka/Ks ratio of 90-6 for the Nigerian isolates compared to the ratio of 14-6 for the Japanese isolates. Most Nigerian strains had the same AhpC amino acid type, indicating that the protein was highly conserved in that population. By contrast, the South African and UK strain sets had ahpC and atpA Ka/Ks ratios of 38-1–65-4, which indicated some variability, with a general trend of decreased variability compared to the Japanese but increased variability compared to the Nigerian set. By contrast, the ureI ratios were of the same order (11-2–16-3) irrespective of geographical origin, but with a bias compared to the other two gene products towards increased non-synonymous substitution and a higher number of UreI amino acid types. The mean ureI ratio of 14-1 for all strain sets in this study closely matched the mean value of 16-2 reported previously by Achtman et al. (1999). As the number of isolates tested was small, the statistical significance of these observations needs to be validated on larger strain sets and their biological significance established.

These associations between markers in H. pylori housekeeping gene sequences and geographical origin were in agreement with other reports that East Asian isolates had independent virulence-associated genotypic features – for instance, the type II cag right-junction motif was present in 95% of East Asian strains compared to 1% or less of
stains from Africa, North America and South Asia (Kersulyte et al., 2000). Moreover, independent cag pathogenicity-island-related evidence that East Asian strains may constitute a genetically distinct population was provided by analysis of the 5’ region of cagA (Yamaoka et al., 1998), of other regions of cagA (van der Ende et al., 1998) and of the right end of the cag pathogenicity island (Kersulyte et al., 2000), as well as by presence of the vacA s1c subtype allele as an East Asian strain marker (van Doorn et al., 1998). Recently, analysis of polymorphism motifs in the ORF (HP0638) encoding the OipA protein segregated East Asian isolates from isolates from the Indian subcontinent and Western countries including African-American isolates (Ando et al., 2002). Our findings support the concept of a nonrandom geographical distribution of certain polymorphism motifs in unrelated genes in the H. pylori genome. No geographical segregation was observed in phylogenetic analysis based either on atpD, scoB, glnA and recA sequences (Solsca et al., 2001) or on glmM sequences (van der Ende et al., 1998), which could indicate that some genes contain only limited or no geographical characteristics although the strain selection in those studies may have influenced interpretation. Sequence disruptions due to frequent recombination and mutation events mask evidence of clonality in all isolates except for those from more recent infections amongst family members (Suerbaum et al., 1998; Owen & Xerry, 2003), and multiple or sequential isolates from the same individual (Falush et al., 2001), so evidence from several unrelated loci supports the relative stability of the geographical motifs. The latter point was confirmed in the present study, as all marker polymorphisms were identical in paired isolates from different gastric sites in 11 of the Japanese and in five of the UK patients and in some family isolates. Overall, the Japanese polymorphisms were relatively rare in H. pylori infecting the UK and South African patients, although data on a larger number of isolates are necessary for rigorous statistical analysis.

Covacci et al. (1999) suggested from the global geographical distribution of H. pylori that the species might have co-evolved with man over prolonged periods of some 50 000 years of human migration leading to genetically distinct human and H. pylori populations. The deduced population structure from multilocus data of the four modern populations of H. pylori suggests that their gene pools were derived from ancestral populations that arose in Africa, Central Asia and East Asia (Falush et al., 2003). However, attempts to define subpopulations within European isolates (the hpEurope population) of H. pylori are confounded by the complex history of human migrations (Falush et al., 2003). Analysis of the UK isolates including the Indian subgroup in the present study showed that while they were distinct from the African and East Asian variants, their gene sequences contained a higher proportion of alleles of African origin. They were also more diverse than the East Asian and Nigerian strains with respect to cagA presence and vacA alleles. Our overall findings are consistent with a human–bacterial comigration hypothesis (Falush et al., 2003) although it is essential for interpretation of population markers to take into account the more recent history of human hosts, as suggested by similarities between Spanish and Latin American strains (Kersulyte et al., 2000). Our investigation of sequences from strains from several neighbouring countries in the East Asian region (Japan, China and Thailand) confirmed the presence of common marker polymorphisms for local geographical variants of H. pylori, and gives support to the concept of a distinct ‘Asian’ clone (Achtman et al., 1999), subsequently designated as the hpEastAsia subpopulation (Falush et al., 2003). The ureI and atpA sequences for Asian clone strains 88-28, R29 and 97-42 included in the present analyses segregated in clade A with our Japanese (Tokyo) isolates, and had the typical East Asian marker polymorphisms. The associations of our European isolates with the hpEurope population represented by strain 26695 await further clarification, as do the associations of the Nigerian strains with the two modern Africa populations (hpAfrica1 and hpAfrica2).

In conclusion, the nucleotide polymorphism analysis described here provides new information on the biogeographical diversity of H. pylori, and in particular identifies specific polymorphisms that could be used to define the Japanese subpopulation, which we propose should be designated H. pylori geovar ‘orientalis’ according to the general criteria of Staley (1999). Application of this biogeographical approach to other population groups could provide a more precise framework for investigating the subtleties of H. pylori–host interactions and disease associations.

ACKNOWLEDGEMENTS

This work was supported in part by the PHLS Central R&D Small Scientific Initiative Fund 2000.

REFERENCES


http://mic.sgmjournals.org 161