Geographical conservation of short inserts in the signal and middle regions of the *Helicobacter pylori* vacuolating cytotoxin gene

Robert J. Owen and Jacqueline Xerry

Campylobacter and Helicobacter Research/Reference Unit, Laboratory of Enteric Pathogens, Centre for Infections, Health Protection Agency, 61 Colindale Avenue, London NW9 5HT, UK

Correspondence

Robert Owen
robert.owen@hpa.org.uk

Short nucleotide sequence inserts within the signal (s) and mid (m) regions of the vacuolating cytotoxin gene (*vacA*) of *Helicobacter pylori* provide the basis for defining the allelic forms widely used for strain typing and as markers for toxin functionality and severity of interactions with host gastric epithelial cells. Here 484 signal region and 411 mid-region sequences (new and from public databases) from 32 countries were analysed to determine the effect of geographical location on insert diversity, which is currently undefined. Short (27 bp) inserts of 52 mol% G + C from 201 sequences (98 %) of the s2 allelic family encoded a highly conserved nine amino acid sequence irrespective of geographical origin. The longer (75 bp) mid-region insert of 38 mol% G + C in 255 sequences of the m2 allelic family was more diverse and represented by 23 peptide variants, with one predominant sequence (MRI type 4) representing 62 % of inserts. Mid-region inserts were widespread throughout European/North American (Western) sequences in the dataset whereas a lower insert frequency was a geographical feature of East Asian sequences. Each insert was preceded by an associated conserved motif that provided a marker of the insertion sites within *vacA*, and facilitated identification of the Chinese m2b genotype. It is concluded that the observed sequence conservation supports the continued global use of *vacA* genotyping, and that inserts could have a functional significance in the mature protein, particularly the s2 form of the toxin, as the same combination of signal and mid-region insert type and preinsert motif was highly conserved.

Received 29 September 2006

Accepted 4 January 2007

INTRODUCTION

The vacuolating cytotoxin (VacA), encoded by the *vacA* gene, is a key pathogenicity factor of *Helicobacter pylori* (Blaser & Atherton, 2004; Gebert *et al.*, 2004), which infects the gastric mucosa of about half of the human population worldwide (Pounder & Ng, 1995). Most infected individuals remain asymptomatic even though the presence of *H. pylori* typically induces chronic gastritis. In some infected individuals (<20 %), *H. pylori* is an important causal factor in the development of peptic ulcer disease (gastric and duodenal ulcers) and gastric low-grade mucosal-associated lymphoid tissue (MALT) lymphoma as well as being a key risk factor in the development of gastric cancer (Suerbaum & Michetti,

2002). The presence of *H. pylori* DNA has been associated with other chronic diseases such as cardiovascular and hepatobiliary disease but viable bacteria have not been isolated from such conditions and direct causality remains unproven (Aceti *et al.*, 2004; Huang *et al.*, 2004; Apostolov *et al.*, 2005).

VacA is a cytotoxin of the gastric epithelial cell layer and a potent immunoregulatory toxin inducing apoptosis in epithelial cells (Boquet *et al.*, 2003; Gebert *et al.*, 2004). The precise mechanisms by which VacA contributes to the onset of chronic disease are unclear but its incidence appears to involve a complex interaction of aetiological factors that include dietary and genetic differences of the host, and bacterial virulence cofactors (Höcker & Hohenberger, 2003). The *vacA* gene, present as a single copy in the genomes of all *H. pylori* tested, is highly polymorphic particularly in the signal and mid-regions (Owen *et al.*, 2002; Blaser & Atherton, 2004). Its mosaic structure is the result of frequent recombination (Atherton *et al.*, 1995; Falush *et al.*, 2001). Strains from many parts of the world have been characterized by the sizes of the regions encoding the N-terminal signal (s) peptide, which determines cytotoxic activity (Letley & Atherton 2000; McClain *et al.*, 2001), and the mid

Abbreviations: MLST, multilocus sequence typing; MR, mid-region; MRI, mid-region insert; MRP, mid-region pre-insert site; SR, signal region; SRI, signal region insert; SRP, signal region pre-insert motif.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are AMO88759–AMO88775.

Accession numbers, source strains and geographical origins of the other GenBank sequences analysed are available as supplementary data with the online version of this paper.

(m) 58 kDa toxin subunit, which includes the cell-binding domain determining target cell specificity (Pagliaccia *et al.*, 1998; Atherton *et al.*, 1999a). Common *vacA* allelic forms (genotypes) identified worldwide are s1/m1 (vacuolating), s1/m2 (selectively vacuolating) and s2/m2 (non-vacuolating), although the latter form is rare in East Asian strains and the s2/m1 form is extremely rare in all populations (Letley *et al.*, 1999). The signal region allelic families can be further divided into s1a, s1b and s1c subfamilies and likewise the mid-region is subdivided into m2a and m2b subfamilies (van Doorn *et al.*, 1998). The *vacA* genotype is of clinical interest as some alleles have been correlated with the level of vacuolating cytotoxin activity in HeLa cells, with the s1/m1 genotype being more toxigenic than the s1/m2 genotype, whereas toxin activity is blocked in the s2/m2 genotype (Letley & Atherton, 2000). The s1 form is often linked to more severe gastric inflammation; even so, the presence of *H. pylori* with a particular *vacA* allelic variant does not necessarily provide a reliable predictor of pathogenic potential (Blaser & Atherton, 2004).

Campbell *et al.* (1997) presented evidence from an investigation of ethnic groups in New Zealand that genetically distinct *H. pylori* lineages may have evolved in parallel with race-specific specialization. Subsequent population genetic studies including multilocus sequence typing (MLST) showed that *H. pylori* from different parts of the world represent seven distinct populations and subpopulations, with modern geographical distributions apparently linked to historical global human migrations (Suerbaum *et al.*, 1998; Achtman *et al.*, 1999; Covacci *et al.*, 1999; Ghose *et al.*, 2002; Falush *et al.*, 2003; Wirth *et al.*, 2004; Kauser *et al.*, 2005). Such genetic analyses have highlighted the panmictic (non-clonal) nature of the species structure and, most significantly, have demonstrated the presence of conserved residual genomic features that can be linked to co-evolution with man. The East Asian (Eastern-type) isolates, particularly those of Japanese and Chinese origin, appear to represent a distinct modern subpopulation. Furthermore, Japanese-specific polymorphisms have been identified within several housekeeping genes (Owen *et al.*, 2004). Geographical separation of *H. pylori* in distinct separately evolving populations is supported by analyses of the gene encoding CagA, a marker of the *cag* pathogenicity island, as well as by *vacA*, in which characteristic differences were identified between isolates originating from China and Japan (East Asian type) compared to those from the USA and Europe (Western type) (van der Ende *et al.*, 1998; Ito *et al.*, 1997; Pan *et al.*, 1998; van Doorn *et al.*, 1998; Yamaoka *et al.*, 1998; van Doorn *et al.*, 1999; Kersulyte *et al.*, 2000). East Asian strains typically have the *vacA* subtype s1c allele, which is rare in other parts of the world (van Doorn *et al.*, 1998). In Europe, a distribution gradient was observed, with most strains in Northern Europe being s1a whereas in Spain and Portugal most strains were s1b. It remains unclear, however, to what extent *H. pylori* strain type might contribute to the four times higher rates of gastric cancer observed in Japanese and other East Asians compared to the United

Kingdom population (Neugut *et al.*, 1996; Lambert *et al.*, 2002).

The application of gene sequence analysis to improve the understanding of population structures of bacterial pathogens (Cooper & Feil, 2004) has resulted in a rapid expansion in the availability of high-quality DNA sequences. As various complete and partial *H. pylori vacA* sequences are now available from public databases, the aim of our study was to utilize that unique information in a comprehensive comparative genomic analysis to define more precisely the global nature of variation of insert sequences within the signal and mid-region allelic families of *vacA* present in the modern *H. pylori* gene pool.

METHODS

Sources of bacterial strains. The 220 isolates of *H. pylori* used for sequencing in-house were selected on the basis of geographical diversity from the reference laboratory collection archived at -80°C . All isolates originated from gastric biopsies of unrelated patients undergoing routine upper gastrointestinal endoscopic investigation for various presentations between 1990 and 2002. The study included Japanese (Tokyo) isolates collected as part of the UK/Japan Gastritis Study from 16 patients (eight males and eight females) aged between 23 and 79 years (mean 55.6 years) described previously (Owen *et al.*, 2004; Naylor *et al.*, 2006). Nine reference cultures of *H. pylori* were obtained in lyophilized form from the National Collection of Type Cultures (NCTC, London): NCTC 11637 (type strain, Australia); NCTC 11638 (Australia); NCTC 13081 (strain Tx30a, USA); NCTC 13082 (Canada), NCTC 13085 (strain 60190, England), NCTC 13086 (Canada), NCTC 11916 (England), NCTC 13207 (CCUG 38772), NCTC 12455 (strain 26695, genome sequenced strain, England; Tomb *et al.*, 1997). Cultures of the genome sequenced strain J99 (USA; Alm *et al.*, 1999) and of strain SS1 (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, England), respectively.

Bacterial culture conditions. All strains of *H. pylori* were recovered from the archived stock of cultures preserved on glass beads at -80°C (Microbank system, Pro-Lab Diagnostics). Strains were sub-cultured for 2–3 days on Columbia agar base (Oxoid) containing 10% (v/v) defibrinated horse blood and incubated at 37°C under microaerophilic conditions (4% oxygen, 5% hydrogen, 5% carbon dioxide and 86% nitrogen) in a MACS-VA500 microaerophilic workstation (Don Whitley Scientific). Identity of *H. pylori* cultures was confirmed by Gram staining and by tests for catalase, cytochrome oxidase and urease activity.

PCR and sequencing. Genomic DNA was extracted from sweep cultures by the method of Wilson (1987). Diluted DNA (100 ng) was used to amplify internal fragments from *vacA* signal- and mid-regions as previously described (Atherton *et al.*, 1999b; Owen *et al.*, 2002). Briefly the primers used for amplification and sequencing were as follows. For the signal region 259 bp fragment, the primer pair used was forward primer VAIF (5'-ATGGAAATACAACAAC-ACAC-3') and reverse primer VAIR (5'-CTGCTTGAATGCGCCAA-AC-3'). For the mid-region 566 bp fragment, the primer pair used was: forward primer VAGF (5'-CAATCTGTCCAATCAAGCGAG-3') and reverse primer VAGR (5'-GCGTCAAAATAATTCCAAGG-3'). The primers were synthesized commercially (MWG Biotech) and PCR amplification reactions were performed in a total volume of 50 μl containing 100 ng diluted template DNA, 1.5 mM MgCl_2 , 0.05 mM of each deoxynucleotide (dATP, dCTP, dGTP and dTTP),

0.4 mM of each oligonucleotide primer, 0.2 µl (1 U) *Taq* polymerase (Invitrogen) and 5 µl 10× buffer, provided by the manufacturer. Automated nucleotide sequencing, from both strands of PCR products, of the internal fragments of *vacA* was then performed using the CEQ dye terminator cycle sequencing quick start kit run on a Beckman CEQ 8000 (Beckman Coulter). The primer sets used enabled complete double-stranded nucleotide sequences to be obtained from a single sequencing run in each direction.

Nucleotide sequences used in comparative genomic analysis. The dataset compiled for this study comprised 550 complete and three partial sequences of *vacA* for isolates of *H. pylori* originating from patients in 32 countries in widely separated geographical regions. The set included a total of 484 *vacA* signal region sequences representing isolates from 30 countries (Table 1). For the signal region s1 allelic family, 60 sequences were determined in-house and 223 were obtained from GenBank. For the signal region s2 allelic family, 159 sequences were determined in-house and 42 were obtained from GenBank. The subset of *vacA* mid-region sequences represented a total of 411 isolates of *H. pylori* from 27 countries (Table 1). For the mid-region m1 allelic family, 43 sequences were determined in-house and 113 were obtained from GenBank. For the mid-region m2 allelic family, sequences of 179 isolates were determined in-house and 76 were obtained from GenBank. Seventeen partial *vacA* (mid-region) sequences determined in this study have been deposited in the EMBL nucleotide sequence database with the accession numbers AMO88759–AMO88775. Accession numbers of the other GenBank sequences used are provided in supplementary Table S1, available with the online version of this paper.

Sequence alignments and phylogenetic analysis. The subsets of *vacA* signal and mid-region sequences (nucleotide and derived amino acid sequences) were analysed independently. Pairwise comparisons and multiple alignments were implemented with BioEdit software (Hall, 1999). Translated and predicted amino acid sequences were analysed visually for marker motifs. Nucleotide and amino acid sequence motifs were compared to public databases using BLAST analysis to identify possible phylogenetic origins. The sets of aligned sequences for signal and mid-regions of *vacA* were analysed by the neighbour-joining method (Saitou & Nei, 1987) using the nucleotide substitution model of Jukes & Cantor (1969). Phylogenetic trees from nucleotide and predicted amino acid sequences rooted with reference to either NCTC 12455 (genome sequenced strain, *vacA* s1/m1 form) or Tx30a (*vacA* s2/m2 form), as well as unrooted trees, were generated in TREECON (van de Peer & de Wachter, 1994).

Statistical analysis. The datasets were analysed using GraphPad InStat version 3.05. Differences between groups were compared using Fisher's exact one-tailed test. Results were considered statistically significant at a *P* value of <0.05.

RESULTS

Polymorphism in the *vacA* signal region inserts

A total of 484 signal region (SR) sequences of *H. pylori vacA* from in-house sequencing and from GenBank were analysed. These comprised 283 sequences each of 259 bp (86 amino acids) representing the *vacA* s1 allele family, and 201 sequences each of 286 bp (95 amino acids) representing the *vacA* s2 allele family. The size of the signal region insert (SRI) that defined the s2 allelic family was consistently 27 bp (9 amino acids) and constituted 9.5 % of the region in all the sequences examined, irrespective of the geographical origin of the strain and of the human hosts' ethnicity.

Table 1. Numbers of *H. pylori vacA* signal and mid-region allele sequences included in dataset by country of origin

Data used were from in-house sequencing and from GenBank (East Asian sequences are shown in bold).

Country	No. of sequences (strains)					
	s1	s2	m1	m2	s-unknown	m-unknown
Argentina	4	–	–	–	–	4
Australia	5	5	4	6	1	1
Bangladesh	9	–	9	2	–	–
Canada	–	10	–	10	–	–
Chile	–	–	1	–	1	–
China	25	–	16	23	22	8 (3)*
China (Hong Kong)	10	–	2	11	6	3
Colombia	13	6	–	–	–	19
Costa Rica	2	1	2	1	–	–
France	–	1	–	1	–	–
Germany	1	–	2	3	4	–
India	9	8	–	3	–	14
Iran	–	–	1	–	1	–
Italy	1	4	1	4	–	–
Japan	55	–	51	3	4	5
Kazakhstan	2	–	–	–	–	2
Kenya	4	1	4	1	–	–
Korea	1	–	–	–	–	1
Mexico	4	–	1	–	–	3
Netherlands	7	3	3	3	2	6
New Zealand	2	–	–	–	–	2
Nigeria	8	–	7	1	–	–
Peru	16	39	11	46	2	–
Poland	–	–	7	–	7	–
Portugal	17	3	6	8	4	10
South Africa	20	19	14	26	–	–
Taiwan	13	–	5	7	10	11
Thailand	6	–	1	4	1	2
Turkey	–	1	–	1	–	–
UK (England)	10	59	6	63	–	–
UK (Wales)	–	16	–	16	–	–
USA	23	17	2	12	1	27
Venezuela	16	8	–	–	–	24
Totals	283	201	156	255	66	142

*Partial sequences.

For the signal region inserts, 98 % had G+C contents of ~52 mol% and nucleotide sequences in which polymorphisms were limited to single synonymous substitutions despite diversity in strain geographical origin, encompassing 17 widely separated countries (Table 1). The nucleotide sequences of each of these inserts translated to give a common predicted nine amino acid sequence NDPIHSESR (SRI type A, Table 2). Inserts with a lower G+C content of 48 mol% were identified in three isolates originating from

Table 2. Features of *H. pylori* signal region (s2 allele) SR inserts

SR insert	Peptide sequence*	No. of isolates†	Reference strain	GenBank no.
Type A	NDPIHSESR	195 (98 %)	Tx30a (USA)	U29401
Type B	NDPIHSENR	3 (1.5 %)	H1906 (UK)	AM088774
Type C	NDPIHSQSR	1 (0.5 %)	H1562 (UK)	AM088775

*The corresponding nucleotide sequences were: type A, AAT-GAT-CCC-ATA-CAC-AGC-GAG-AGT-CGC; type B, AAT-GAT-CCC-ATA-CAC-AGC-GAG-AAT-CGC; and type C, AAC-GAT-CCC-ATA-CAC-AGC-CAG-AGT-CGC (key single nucleotide polymorphisms are underlined in bold). Defined by reference to locations within the sequences of the *vacA* signal protein coding regions in genotype s2/m2 strain Tx30a (U29401).

†In total, 201 sequences for the s2 allele were analysed but two sequences were excluded from the above analysis because they were mixed and could not be fully resolved.

Wales (one patient) and from South Africa (two patients); these had a single S to N substitution at amino acid position 8 (SRI type B). The insert sequence of a strain from a patient in England (London) had a markedly higher G + C content of 56 mol% and contained a single E to Q substitution at amino acid position 7 (SRI type C). No homologies between any of these insert sequences and other known nucleotide or amino acid sequences were identified in a BLAST search.

Polymorphism in *vacA* signal regions adjacent to the insert

Overall diversity of s1 alleles. The 484 complete *vacA* signal region sequences of both the genotype s1 and s2 allelic subfamilies included in the analysis were highly polymorphic, with 345 nucleotide sequence types (127 amino acid sequence types), defined on the presence of one or more nucleotide substitutions when compared to the reference sequence. The 259 bp amplicons of the 283 isolates in the s1 allelic subfamily had conserved G + C contents (48–49 mol%, mean 49 mol%) irrespective of geographical origin but were heterogeneous at the nucleotide level and were characterized by 237 nucleotide sequence types (85 amino acid sequence types). Nevertheless, nearest neighbour analysis of these sequences (nucleotide and predicted amino acid) indicated a high degree of overall relatedness (>75 %) with no significant branches or subgroups within the s1 subset of sequences irrespective of the geographical origin of the isolates.

Overall diversity of s2 alleles. The 286 bp amplicons from 201 isolates of the s2 allelic subfamily had conserved G + C contents (47–51 mol%, mean 50 mol%) and, like the s1 allele family sequences, were heterogeneous at the nucleotide level, with 108 nucleotide sequence types (42 amino acid sequence types). Nearest neighbour analysis of these sequences (nucleotide and predicted amino acid) indicated a high degree of overall relatedness (>75 %) between sequences with no significant branches or subgroups irrespective of the geographical origin of the isolates. Because of the rarity of the s2 allele family in the

East Asian populations, no Japanese or Chinese sequences were available for inclusion in the analyses.

Features of pre-insert site sequences. Alignments of s-region sequences showed a conserved amino acid motif within the signal region immediately adjacent to the 5'-end of the insert. Most (87 %) s2 sequences contained a signal region pre-insert motif (SRP type I) with ten amino acids giving the sequence MGTELGANTP immediately before the insert site (Fig. 1a). BLAST homology searches confirmed that the SRP motif was not present elsewhere in the *H. pylori* genome or in any other bacterial sequences in GenBank. Twenty-five other s2 strains were represented by five other SRP types (II to VI) defined on single or double substitutions within the motif. For instance, a further 10 % of isolates had either M(1) to I or G(2) to S substitutions. By contrast, when the insert was not present in *vacA*, 68 % of s1 allele strains had the completely different VSITPQQSHA (35 %) or VSITPQKSHA (33 %) sequence at the SRP motif location, with 88 other strains represented by 11 variants containing minor amino acid substitutions.

Results summarized in Table 3 show that the signal region pre-insert motif (SRP) type I, present in strains such as Tx30a, was more commonly associated with the presence of signal region insert (SRI) type A than with other insert types (82 % vs 18 %). Thus the two groups of strains defined on the basis of presence or absence of insert type A differed significantly ($P=0.001$) even though the combination of signal region insert type A with pre-insert motif type I was a feature of geographically diverse strains.

Polymorphism in the *vacA* mid-region inserts

A set of 408 complete sequences for the *H. pylori vacA* mid-region (MR), an internal region that included the ORF encoding the p58 toxin subunit (cell binding domain) was analysed using multiple alignments (the three incomplete sequences were excluded from such analyses). These sequences included 255 MR sequences of 641 bp (213 amino acids) for strains of the m2 allelic subfamily. The size of the

Fig. 1. Locations of inserts and associated insert site motifs (labelled rectangles) in *H. pylori vacA* translated amino acid sequences. Unlabelled rectangles indicate other putative conserved motifs under further analysis. Upper sequence is the reference amino acid sequence used in each alignment (corresponding amino acid number is given at start of each sequence): (a) Signal region. NCTC 12455 (AE000598) for s1 allele (upper sequence) vs Tx30a (U29401) for s2 allele (lower sequence); (b) Mid-region. NCTC 12455 for m1 allele (upper sequence) vs Tx30a for m2 allele (lower sequence); (c) Mid-region. Tx30a for m2-A allele subfamily variant (upper sequence) vs CHN3295b (AF050319) for novel m2-B allele subfamily variant (lower sequence). In the sequence alignments, amino acid identities are shown by dashes and insertions/deletions by a continuous line.

mid-region insert (MRI) located at nucleotide position 1560 (amino acid position 520 defined relative to the Tx30a *vacA* open reading frame) was highly conserved at 75 bp (25 amino acids) and constituted 12 % of the mid-region. The G + C contents of the inserts were also highly conserved and were typically 37 to 40 mol% (mean 38 mol%) except for 15 inserts that were mostly of East Asian origin with lower G + C contents of 34–35 mol%.

Diversity amongst the mid-region insert peptide sequences was evident when they were aligned with the insert sequence of strain Tx30a, the well-characterized non-toxicogenic strain (McClain *et al.*, 2001) used here as the reference for the *vacA* m2-allelic subfamily. The Tx30a mid-region insert with the amino acid motif LRVNGHTAHFKNIDATKSDNGLNTS was designated the reference for MRI type 1. A total of 23 amino acid variants (arbitrarily numbered 1 to 23) were defined with each differing by multiple amino acid substitutions when compared with the MRI type 1 sequence. The frequencies of the predominant MRI sequence types (all other types were represented by just one or two sequences) are given in Table 4 with the amino acid substitutions used for defining each type. The commonest form was MRI type 4, which was a feature of 158 (62 %) geographically diverse strains of the m2 allelic subfamily, and was defined by the 25 amino acid sequence LRVNGHSAHFKNIDASKSDNGLNTS. The second most common variant was MRI type 2 with the 25 amino acid sequence LRVNGHSAHFKNIDATKSDNGLNTS, which was a feature of 37 (14 %) geographically diverse strains. The type 4 and type 2 forms differed only in a single S to T substitution at amino acid 16 and together represented 76 % of all available MR insert sequences (Table 4). Five other less common (up to 6 %) MRI forms were also characterized by single amino acid substitutions when compared to the type 1 sequence. Most MR inserts characteristically contained a short internal amino acid SDNGLN motif (positions 18 to 23) except for inserts from 15 MRI type 22 sequences (from China and Peru) that were mainly of Chinese origin. These all contained a unique GRNGID sequence that was also found in MRIs from several other miscellaneous Chinese sequences that were not assigned an MRI type.

Polymorphism in the *vacA* mid-region adjacent to the insert

Overall diversity of m1 alleles. The 156 *vacA* sequences of 566 bp (188 amino acids) for the m1 allelic family had highly conserved G + C contents (39–40 mol%, mean 40 mol%) but were heterogeneous at the nucleotide level, with numerous single nucleotide polymorphisms giving 146 nucleotide sequence types (109 deduced amino acid sequence types). Nearest neighbour analysis of these mid-region m1 sequences (nucleotide and predicted amino acid sequences) nevertheless confirmed a high (>75 %) degree of overall relatedness between sequences, irrespective of geographical origin. There were no distinctions between the m1 alleles representing European/North American/African sequences and the 19 sequences of East Asian origin.

Table 3. Associations between predicted amino acid sequences of predominant inserts and corresponding pre-insert site sequences in *vacA* of *H. pylori*

For the locations of the motifs in *VacA*, see Fig. 1.

<i>vacA</i> allele	Insert sequence type	Pre-insert site sequence type	Isolates with both types	<i>P</i> value	Reference strain (GenBank sequence), country of isolation and <i>vacA</i> genotype
S2	SRI type A	SRP type I	164	<0.001	Tx30a (U29401), USA, s2m2
M2	MRI type 4	MRP type 1	159	<0.001	H1906 (AM088774), UK (Wales), s2m2
M2	MRI type 2	MRP type 1	36	0.011	H5451 (AM088766), Canada, s2m2
M2	MRI type 22	MRP type 3	14	0.004	HK-41 (AJ390658), China (Hong Kong), s1m2

Overall, however, our analyses showed the m1 sequences were phylogenetically distinct from those of the m2 alleles based on comparison with the Tx30a reference sequence, as reported previously by van Doorn *et al.* (1999).

Overall diversity of *vacA* m2 alleles. The 255 *vacA* sequences of 641 bp (213 amino acids) of the m2-allelic subfamily strains likewise all had highly conserved G+C contents (40–42 mol%) and were also heterogeneous at the nucleotide level, with 179 nucleotide sequence variants (155 deduced amino acid sequence variants). Nearest neighbour analysis of these mid-region m2 allele sequences (nucleotide and predicted amino acid sequences) indicated a high degree of overall sequence relatedness (>75%), irrespective of geographical origin. As with the m1 sequences, there were no clear-cut distinctions between European/North American/African sequences. However, a subset of 14 Chinese sequences formed a distinct separate lineage that included three sequences (HK-41, HK-46 and CH-4), which was consistent with the previously reported subdivision of m2 alleles into m2a and m2b genotypes (van Doorn *et al.*, 1998).

Features of pre-insert site sequences. Multiple alignments of the 255 complete *vacA* sequences of the mid-region m2 genotype showed a short conserved motif that was located immediately upstream of the insert and was

designated the mid-region pre-insert site (MRP) motif. That 11 amino acid motif (MRP type 1) had the sequence NGNIYLGKSTN (Fig. 1b) and was a feature of 215 sequences (84%) representing the m2 alleles including strainTx30a (m2a genotype). A variant (MRP type 2) with an I to V substitution at motif position 4 was found in 23 strains (9%), while another variant (MRP type 3) with four amino acid substitutions (positions 4, 8, 9 and 11) was a feature of 16 Chinese sequences including the genotype m2b sequences described by van Doorn *et al.* (1998). One additional sequence (MRP type 4) from a Peruvian strain resembled the latter type except for an M to L substitution at position 4.

By contrast, when the insert was absent, 91 sequences representing the m1 genotype including those of NCTC 12445, NCTC 11637 and J99 had the KGIDTGNNGFN sequence at that position. A similar sequence was present in 59 Asian m1 strains but there was a G to D substitution at position 2. Likewise, minor amino acid variations were present in five other m1 genotype strain sequences. Different conserved motifs identified at additional locations within the m-region (see unlabelled boxes in Fig. 1) will be described elsewhere.

Results summarized in Table 3 show that the two most common inserts MRI types (2 and 4) and the respective

Table 4. Amino acid substitutions defining the most common variants of mid-region insert (m2) sequences (MRIs) of geographically diverse *H. pylori vacA*

MRI type	Frequency	Peptide sequence motif*	Comments
4	158 (62%)	LRVNG HSAHF KNIDA <u>SKSD</u> NGLNTS	Widespread
2	37 (14%)	LRVNG HSAHF KNIDA TKSD NGLNTS	Widespread
22	15 (6%)	LKVNG HTANF KDIDA <u>SKGR</u> NGIDT <u>T</u>	China/Peru
9	9 (4%)	KRVNG HSAHF KNIDA TKSD NGLNTS	China/Germany/Japan/Peru/UK/USA
14	6 (2%)	LRVNG HSAHF KNIDA <u>SKSD</u> NGLNT <u>T</u>	Australia/Canada/Portugal/UK/USA
16	5 (2%)	LKVNG HSAHF KNIDA <u>SKSDN</u> GLNTS	China/Netherlands/Peru/Taiwan
20	3 (1%)	LKVNA NTAHF KDIDA <u>SKSDN</u> GLNTS	South Africa

*Defined by reference to locations within the *vacA* sequence of s2/m2 strain Tx30a (U29401; MRI type 1). Type-specific polymorphisms compared to the reference sequence are in underlined bold.

mid-region pre-insert motif (MRP type 1) were significantly associated ($P=0.0001$). These types represented isolates from many countries so the associations were independent of geographical origin. By contrast, a third insert type, MRI type 22, which was a feature of strains of Chinese origin, was significantly associated with pre-insert motif sequence MRP type 3 ($P=0.004$). Overall, the analysis showed that 119 strains, representing 71 % of the available complete s2/m2 allele sequences, had the insert/motif combination SRI type A/MRI-type 4/MRP type 1.

DISCUSSION

The availability of extensive sequence data on *vacA* offers a unique opportunity to explore the effects of human geographical separation on genomic variation within that locus, which is considered to be a key factor in *H. pylori*-host interactions (Gebert *et al.*, 2004; Cover & Blanke, 2005). The aim of the present study was to analyse diversity amongst the short *vacA* insert sequences using new sequences and all currently available online database sequences. Although the presence or absence of *vacA* inserts provides the basis of a widely used scheme for *H. pylori* strain genotyping (Atherton *et al.*, 1999b), the clinical significance of a particular genotype remains unclear, and consequently *vacA* typing is of limited value in predicting severity of infection and possible clinical outcome (Go *et al.*, 1998; Blaser, 1999; Höcker & Hohenberger, 2003). Analyses of the 892 sequences of the *vacA* s and m alleles in the dataset compiled for this study demonstrated that insert sequences were highly conserved according to a number of criteria, notably their location within *vacA*, overall nucleotide length, G + C content and similarities within allelic families in nucleotide and hence in predicted amino acid sequences. Extensive recombination is a recognized feature of the *H. pylori* genome within both housekeeping and pathogenicity-associated genes (Suerbaum *et al.*, 1998). Furthermore, there is evidence of gain and loss of multiple genes during evolution of the species (Gressmann *et al.*, 2005). The globally conserved nature of these short *vacA* inserts evident from our analysis was therefore unexpected. Despite the presence of numerous nucleotide polymorphisms in the sets of signal- and mid-region alleles, we found that almost all (98 %) of the signal region inserts were defined by the unique NDPIHSESR amino acid sequence. Likewise, 76 % of the mid-region inserts in the sequence dataset also were highly conserved at the predicted amino acid level, as most nucleotide mutations were synonymous, irrespective of geographical origin of the *H. pylori* isolates. The predominant m-insert sequences were represented by just two peptide motifs (types 2 and 4), which differed only by a single serine to threonine polymorphism.

The mid-region and signal-region peptide sequences were unrelated, and the insert sequences shared no homologies with either of the two sequenced genomes of *H. pylori* (Tomb *et al.*, 1997; Alm *et al.*, 1999), with the *Helicobacter hepaticus* genome (Suerbaum *et al.*, 2003), or with genomes

of *Campylobacter jejuni* and other allied enteric bacterial species deposited in GenBank. The possible origins of the phylogenetically unrelated inserts in *H. pylori vacA* therefore remain unclear. Their sequence and size uniqueness, and global distribution combined with a high degree of conservation suggest they may have ancestral origins rather than representing recent acquisitions by the *H. pylori* genome. Although we have referred to the target sequences as 'inserts' for the purpose of this study, they could possibly be viewed as ancestral remnants of regions that have undergone selective deletion from *vacA* leading to the emergence of more pathogenic forms of *H. pylori* during adaptation to colonize the human gastric mucosa in different geographical environments. During human migrations over millennia, there could have been selective loss of forms containing the insert, resulting in the eventual emergence of the more virulent *vacA* s1 form as the predominant component of local species gene pools. Interestingly, there is evidence suggesting a link between the presence of *vacA* s2 alleles and the absence of *cagA* (Andreson *et al.*, 2002) and by inference possibly the complete *cag* pathogenicity island. Such an association could be interpreted as an indicator of the primitive origins of the s-insert, although it can only be speculated that ancestral forms of *H. pylori* may have lacked the *cag* pathogenicity island (Gressmann *et al.*, 2005).

A key feature of the global epidemiology of *vacA* is the non-random nature of the frequencies of the different allelic forms. Surveillance data on naturally occurring *H. pylori* strain populations show consistent differences in the relative frequencies of the two allelic families; for instance, in England and Wales up to 65 % of m alleles, depending on the demographics of patients sampled, had an insert whereas 11 % or fewer have s alleles containing an insert (Owen *et al.*, 2002; Elviss *et al.*, 2004, 2005). Likewise, similar frequencies of the s2 and m2 genotypes have been reported in many other parts of the world (van Doorn *et al.*, 1999). Overall, these cumulative data highlight the fact that the s1 allelic family is the predominant *vacA* signal form of strains infecting humans globally with suballelic forms showing some geographical associations; for instance, allele s1c was described as the major subtype in East Asian strains whereas allele s1a was more typical of Western, in particular Northern European, strains (van Doorn *et al.*, 1999; Yakoob *et al.*, 2002; Wang *et al.*, 2003; Zhou *et al.*, 2004a, b). It should be noted that these subtypes were determined by differences in the main signal region rather than within the insert. No subtypes within s2 forms have been defined, a fact that reflects the rarity and conserved nature of such inserts in present *H. pylori* populations globally. For the *vacA* m alleles, the m1 and m2a forms generally have equal frequencies, except in Spain and Central and South America, where m1 is more prevalent, and in China, where the m2a form predominates (van Doorn *et al.*, 1999; Wang *et al.*, 2003).

The relative frequencies of the combined s/m allelic forms are also highly conserved but some differences according to geographical location have been noted; for instance, subtype

m2 was found exclusively amongst East Asian s1c strains (van Doorn *et al.*, 1999). The results on isolates from English and Welsh dyspeptics (Owen *et al.*, 2002; Elviss *et al.*, 2004, 2005) were consistent with data from other parts of the world in showing that the *vacA* s1/m1 and s1/m2 genotypes occur at similar frequencies and together constitute up to 90 % of local *H. pylori* gene pools. The frequency of forms with inserts in both the signal and mid-regions (s2/m2 form), by contrast, was lower at about 15 % or less, while isolates with an insert in the signal but not the mid-region (s2/m1 form) were rare and only found in South Africa (Letley *et al.*, 1999). The s2 alleles occur only in combination with m2 alleles, which suggests that the s2/m1 combination results in a disabled non-functional form of VacA toxin, and consequently that such strains would be unable to colonize and/or persist on gastric epithelia. Data on human populations in European, African and North American countries as well as in China show that the frequency of the s2/m2 allelic family is low (generally around 10 %). In Japanese human populations, the s2/m2 form appears even rarer: for example, such forms were found in only 4 % of *H. pylori* infecting patients in Okinawa (Zhou *et al.*, 2004b) and were completely absent from isolates infecting adult gastritis patients in Tokyo (Owen *et al.*, 2004). In Japanese children the predominant genotype reported was s1c/m1b (Azuma *et al.*, 2004). There may well be differences in allelic frequencies within the numerically vast and geographically diverse East Asian population. The s1/m2 form is prevalent in China although there are indications of regional differences (Pan *et al.*, 1998; Yakoob *et al.* 2002; Zhou *et al.*, 2004a, b; Gong *et al.*, 2005). For instance, in the Shaanxi Province (Central China) the relative frequencies for the s1/m1, s1/m2 and s2/m2 allele combinations were 52 %, 39 % and 9 % respectively (Qiao *et al.*, 2003), whereas in two areas of Northern China, the predominant genotypes were s1/m1b and s1/m2 (no s2 forms were reported), in individuals with multiple strain infections (Gong *et al.*, 2005). Overall these results highlight the complexity of *H. pylori* strain populations and the difficulty of drawing general conclusions.

It has been proposed that VacA may contribute to colonization and persistence *in vivo* at different stages of the infection process as part of its multifunctional role (Cover & Blanke, 2005). The fact our results indicate a high degree of conservation within the *vacA* inserts suggests that other factors may also contribute to overall VacA functionality. The simple presence or absence of the insert appears to be the key functional marker in the case of the signal region. Here the direct effect of the presence of the insert is to cause an amino acid terminal (hydrophilic) extension of VacA that in turn either causes blocking or markedly alters the functional properties of the toxin, leading to loss of toxigenic activity (Letley & Atherton, 2000; McClain *et al.*, 2001). The VacA proteins encoded by the s2/m2 forms nevertheless apparently retain an important biological function in some individuals (McClain *et al.*, 2001); for instance, s2/m2 allele strains from cases of chronic gastric disease in England have proved difficult to eradicate despite

repeated antibiotic therapy. Contributing factors may be intrinsic antibiotic resistance as well as lower metabolic activity and low colonization levels of s2 strains. As a consequence they may be more difficult to detect with diagnostic tests such as the urea breath test. By contrast, the mid-region insert appears to have no role in determining target cell specificity, with differences between the m1 and m2 forms of the protein being determined by a short region of amino acids from the beginning of the mid-region to the m1–m2 junction (Ji *et al.*, 2000).

The multiple alignments we performed of complete *vacA* sequences showed two important features in relation to geographical diversity of alleles. Firstly, there were high levels of homology within geographically diverse sequences defined as belonging to either the signal or mid-region families of alleles. Our phylogenetic analysis confirmed the close relatedness reported previously amongst the m1 allelic family (Ito *et al.*, 1997) and the more marked distinction between the m1 and m2 forms (Pagliaccia *et al.*, 1998; Atherton *et al.*, 1999a; Ji *et al.*, 2002). Our analysis also confirmed the findings of van Doorn *et al.* (1998) of divergence within the m2 allelic family with two distinct subfamilies of sequences, one of which was predominantly of Chinese origin (designated here as m2 B). Although the level of similarity between these two m2 subgroups is unclear, the uniqueness of the m2 B subgroup was attributable to polymorphisms in the m-insert, all of which had closely related forms (MRI types 21 and 22), as well as at other locations within the m-region (Fig. 1c). These results further support the existence of two possible *vacA* m2 sequence sublineages within the Chinese *H. pylori* gene pool. Sequences of *vacA* appear to highlight finer differences in population structure than discernible from MLST analyses based on housekeeping genes, which grouped Chinese and Japanese isolates in one East Asian subpopulation designated hspEAsia (Suerbaum & Achtman, 2004). The distinctiveness of the Chinese *H. pylori* gene pool is also evident from analyses of *cagA*, which suggest a more active East Asian form (Zhou *et al.*, 2004b; Zhang *et al.*, 2005), and from analysis of *hopQ* alleles (Cao *et al.*, 2005).

The second interesting finding in our analysis was the presence of several conserved short sequences in core *vacA* regions that provided markers for a specific location directly associated with the gaining of an insert. The SRP motif identified upstream of the signal-region insert provided a flanking marker in 92 % of strains for the location of the signal region insert. Likewise the MRP motif identified upstream of the mid-region insert provided a flanking marker in 84 % of strains for the location of the 75 bp mid-region insert. The locations of these motifs fell within the region identified as determining the phenotypic differences between the m1 and m2 forms of the proteins in target cell specificity (Ji *et al.*, 2000). The characterization of SRP and MRP motifs provides the first evidence of site-specific markers within *vacA* for the insert sequences that might explain the constancy of their locations, despite an

enormous geographical diversity of isolate origin. This has a functional relevance as the SRP motif contains the signal-region sequence processing site, which differs in the s1 and s2 forms, and as a consequence, the extension of the mature N-terminus is three amino acids less as a result of the insert. In view of the role of this extension in blocking vacuolating activity (Letley & Atherton 2000; Letley *et al.*, 2003; McClain *et al.*, 2001), it is an important finding that this processing site is well conserved amongst s2 strains as it suggests similar signal processing and, together with the insertion, a conserved N-terminal extension to the s2 VacA form.

In conclusion, our comprehensive analysis of geographically diverse *vacA* sequences demonstrated that the signal and mid-region inserts providing the basis for *vacA* allelic typing have several unique conserved features. We speculate that internal insert variation appears unlikely to contribute markedly to differences in *H. pylori* strain virulence but nevertheless the findings support the continued use of the insertion-typing primers for determining *vacA* type. Future work should determine how geographical variation at other locations within *vacA* might determine variation in toxin structure and function as these could provide more specific markers for use as predictors of severity of *H. pylori*-host interactions. The identification of a *H. pylori* subtype found only in Chinese patients highlights the need to extend investigations into geographical diversity, particularly in regions where there is a high incidence of gastric cancer and host factors may favour colonization by particular strain pathotypes.

REFERENCES

- Aceti, A., Are, R., Sabino, G., Fenu, L., Pasquazzi, C., Quaranta, G., Zechini, B. & Terrosu, P. (2004). *Helicobacter pylori* active infection in patients with acute coronary heart disease. *J Infect* **49**, 8–12.
- Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z. J., Suerbaum, S., Thompson, S. A., van der Ende, A. & van Doorn, L. J. (1999). Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol Microbiol* **32**, 459–470.
- Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C. & other authors (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180.
- Andreson, H., Lõivukene, K., Sillakivi, T., Maarsoos, H. I., Ustav, M., Peetsla, A. & Mikelsaar, M. (2002). Association of *cagA* and *vacA* genotypes of *Helicobacter pylori* with gastric diseases in Estonia. *J Clin Microbiol* **40**, 298–300.
- Apostolov, E., Al-Soud, W. A., Nilsson, I., Kornilovska, I., Usenko, V., Lyzoqubov, V., Gaydar, Y., Wadstron, T. & Ljungh, A. (2005). *Helicobacter pylori* and other *Helicobacter* species in gall bladder and liver of patients with chronic cholecystitis detected by immunological and molecular methods. *Scand J Gastroenterol* **40**, 96–102.
- Atherton, J. C., Cao, P., Peek, R. M., Jr, Tummuru, M. K., Blaser, M. J. & Cover, T. L. (1995). Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. Association of specific *vacA* types with cytotoxin production and peptic ulceration. *J Biol Chem* **270**, 17771–17777.
- Atherton, J. C., Sharp, P. M. & Cover, T. L. (1999a). Vacuolating cytotoxin (*vacA*) alleles of *Helicobacter pylori* comprise two geographically widespread types, m1 and m2, and have evolved through limited recombination. *Curr Microbiol* **39**, 211–218.
- Atherton, J. C., Cover, T. L., Twells, R. J., Hawkey, C. J. & Blaser, M. J. (1999b). Simple and accurate PCR-based system for typing vacuolating cytotoxin alleles of *Helicobacter pylori*. *J Clin Microbiol* **37**, 2979–2982.
- Azuma, T., Kato, S., Zhou, W., Yamazaki, S., Yamakawa, A., Ohtani, M., Fujiwara, S., Minoura, T., Iinuma, K. & Kato, T. (2004). Diversity of *vacA* and *cagA* genes of *Helicobacter pylori* in Japanese children. *Aliment Pharmacol Ther* **20**, 7–12.
- Blaser, M. J. (1999). Allelic variation in *Helicobacter pylori*: progress but no panacea. *Gut* **45**, 499–502.
- Blaser, M. J. & Atherton, J. C. (2004). *Helicobacter pylori* persistence: biology and disease. *J Clin Invest* **113**, 321–333.
- Boquet, P., Ricci, V., Galmiche, A. & Gauthier, N. C. (2003). Gastric cell apoptosis and *Helicobacter pylori*: has the main function of *vacA* finally been identified? *Trends Microbiol* **11**, 410–413.
- Campbell, S., Fraser, A., Holliss, B., Schmid, J. & O'Toole, P. W. (1997). Evidence for ethnic tropism of *Helicobacter pylori*. *Infect Immun* **65**, 3708–3712.
- Cao, P., Lee, K. J., Blaser, M. J. & Cover, T. L. (2005). Analysis of *hopQ* alleles in East Asian and Western strains of *Helicobacter pylori*. *FEMS Microbiol Lett* **251**, 37–43.
- Cooper, J. E. & Feil, E. J. (2004). Multilocus sequence typing – what is resolved? *Trends Microbiol* **12**, 373–377.
- Covacci, A., Telford, J. L., Del Giudice, G., Parsonnet, J. & Rappuoli, R. (1999). *Helicobacter pylori* virulence and genetic geography. *Science* **284**, 1328–1333.
- Cover, T. L. & Blanke, S. R. (2005). *Helicobacter pylori vacA*, a paradigm for toxin multifunctionality. *Nat Rev Microbiol* **3**, 320–332.
- Elviss, N. C., Owen, R. J., Xerry, J., Walker, A. M. & Davies, K. (2004). *Helicobacter pylori* antibiotic resistance patterns and genotypes in adult dyspeptic patients from a regional population in North Wales. *J Antimicrob Chemother* **54**, 435–440.
- Elviss, N. C., Owen, R. J., Breathnach, A., Palmer, C. & Shetty, N. (2005). *Helicobacter pylori* antibiotic-resistance patterns and risk factors in adult dyspeptic patients from ethnically diverse population in Central and South London during 2000. *J Med Microbiol* **54**, 567–574.
- Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M. & Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A* **98**, 15056–15061.
- Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., Blaser, M. J., Graham, D. Y., Vacher, S. & other authors (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585.
- Gebert, B., Fischer, W. & Haas, R. (2004). The *Helicobacter pylori* vacuolating cytotoxin: from cellular vacuolation to immunosuppressive activities. *Rev Physiol Biochem Pharmacol* **152**, 205–220.
- Ghose, C., Perez-Perez, G. I., Dominguez-Bello, M. G., Pride, D. T., Bravi, C. M. & Blaser, M. J. (2002). East Asian genotypes of *Helicobacter pylori* strains in Amerindians provide evidence for its ancient human carriage. *Proc Natl Acad Sci U S A* **99**, 15107–15111.
- Go, M. F., Cissell, L. & Graham, D. Y. (1998). Failure to confirm association of *vacA* gene mosaicism with duodenal ulcer disease. *Scand J Gastroenterol* **33**, 132–136.
- Gong, Y. H., Wang, Y. & Yuan, Y. (2005). Distribution of *Helicobacter pylori* in North China. *World J Gastroenterol* **11**, 3523–3527.

- Gressmann, H., Linz, B., Ghai, R., Pleissner, K. P., Schlappbach, R., Yamaoka, Y., Kraft, C., Suerbaum, S., Meyer, T. F. & Achtman, M. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genetics* **1**, 419–427.
- Hall, T. A. (1999). BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.
- Höcker, M. & Hohenberger, P. (2003). *Helicobacter pylori* virulence factors – one part of a big picture. *Lancet* **362**, 1231–1233.
- Huang, Y., Fan, X. G., Wang, Z. M., Zhou, J. H., Tian, X. F. & Li, N. (2004). Identification of *Helicobacter* species in human liver samples from patients with primary hepatocellular carcinoma. *J Clin Pathol* **57**, 1273–1277.
- Ito, Y., Azuma, T., Ito, S., Miyaji H., Hirai, M., Yamazaki, Y., Sato, F., Kato, T., Kohli, Y. & Kuriyama, M. (1997). Analysis and typing of the *vacA* gene from *cagA*-positive strains of *Helicobacter pylori* isolated in Japan. *J Clin Microbiol* **35**, 1710–1714.
- Ji, X., Fernandez, T., Burrone, D., Pagliaccia, C., Atherton, J. C., Reyat, J. M., Rappuoli, R. & Telford, J. L. (2000). Cell specificity of *Helicobacter pylori* cytotoxin is determined by a short region in the polymorphic midregion. *Infect Immun* **68**, 3754–3757.
- Ji, X., Frati, F., Barone, S., Pagliaccia, C., Burrone, D., Xu, G., Rappuoli, R., Reyat, J. M. & Telford, J. L. (2002). Evolution of functional polymorphism in the gene coding for the *Helicobacter pylori* cytotoxin. *FEMS Microbiol Lett* **206**, 253–258.
- Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism III*, pp. 21–32. edited by H. N. Munro. New York: Academic Press.
- Kauser, F., Hussain, M. A., Ahmed, I., Srinivas, S., Devi, S. M., Majeed, A. A., Rao, K. R., Khan, A. A., Sechi, L. A. & Ahmed, N. (2005). Comparative genomics of *Helicobacter pylori* isolates recovered from ulcer disease patients in England. *BMC Microbiol* **5**, 32.
- Kersulyte, D., Mukhopadhyay, A. K., Velapatiño, B., Su, W., Pan, Z., Garcia, C., Hernandez, V., Valdez, Y., Mistry, R. S. & other authors (2000). Differences in genotypes of *Helicobacter pylori* from different human populations. *J Bacteriol* **182**, 3210–3218.
- Lambert, R., Guilloux, A., Oshima, A., Pompe-Kirn, V., Bray, F., Parkin, M., Ajiki, W. & Tsukuma, H. (2002). Incidence and mortality from stomach cancer in Japan, Slovenia and the USA. *Int J Cancer* **97**, 811–818.
- Lee, A., O'Rourke, J., De Ungria, M. C., Robertson, B., Daskalopoulos, G. & Dixon, M. F. (1997). A standardized mouse model of *Helicobacter pylori* infection: introducing the Sydney strain. *Gastroenterology* **112**, 1386–1397.
- Letley, D. P. & Atherton, J. C. (2000). Natural diversity in the N terminus of the mature vacuolating cytotoxin of *Helicobacter pylori* determines cytotoxin activity. *J Bacteriol* **182**, 3278–3280.
- Letley, D. P., Lastovica, A., Louw, J. A., Hawkey, C. J. & Atherton, J. C. (1999). Allelic diversity of the *Helicobacter pylori* vacuolating cytotoxin gene in South Africa: rarity of the *vacA* s1a genotype and natural occurrence of an s2/m1 allele. *J Clin Microbiol* **37**, 1203–1205.
- Letley, D. P., Rhead, J. L., Twells, R. J., Dove, B. & Atherton, J. C. (2003). Determinants of non-toxicity in the gastric pathogen *Helicobacter pylori*. *J. Biol Chem* **278**, 26734–26741.
- McClain, M. S., Cao, P., Iwamoto, H., Vinion-Dubiel, A. D., Szabo, G., Shao, Z. & Cover, T. L. (2001). A 12-amino-acid segment, present in type s2 but not type s1 *Helicobacter pylori* VacA proteins, abolishes cytotoxin activity and alters membrane channel formation. *J Bacteriol* **183**, 6499–6508.
- Naylor, G. M., Gotoda, T., Dixon, M., Shimoda, T., Gatta, L., Owen, R., Tompkins, D. & Axon, A. (2006). Why does Japan have a high incidence of gastric cancer? Comparison of gastritis between UK and Japanese patients. *Gut* **55**, 1545–1552.
- Neugut, A. I., Hayek, H. & Howe, G. (1996). Epidemiology of gastric cancer. *Semin Oncol* **23**, 281–291.
- Owen, R. J., Xerry, J., Peters, T. M. & Teare, E. L. (2002). Surveillance and clinical relevance of *vacA* genotypes of *Helicobacter pylori* infecting dyspeptic patients in mid Essex. *Commun Dis Public Health* **5**, 106–111.
- Owen, R. J., Xerry, J., Gotada, T., Naylor, G. & Tompkins, D. (2004). Analysis of geospecific markers for *Helicobacter pylori* variants in patients from Japan and Nigeria by triple-locus nucleotide sequence typing. *Microbiology* **150**, 151–161.
- Pagliaccia, C., de Bernard, M., Lupetti, P., Ji, X., Burrone, D., Cover, T. L., Papinin, E., Rappuoli, R., Telford, J. L. & Reyat, J. M. (1998). The m2 form of the *Helicobacter pylori* cytotoxin has cell type-specific vacuolating activity. *Proc Natl Acad Sci U S A* **95**, 10212–10217.
- Pan, Z. J., Berg, D. E., van der Hulst, R. W., Su, W. W., Raudoukiene, A., Xiao, S. D., Dankert, J., Tygat, G. N. & van der Ende, A. (1998). Prevalence of vacuolating cytotoxin production and distribution of distinct *vacA* alleles in *Helicobacter pylori* from China. *J Infect Dis* **178**, 220–226.
- Pounder, R. E. & Ng, D. (1995). The prevalence of *Helicobacter pylori* infection in different countries. *Aliment Pharmacol Ther* **9**, 33–39.
- Qiao, W., Hu, J. L., Xiao, B., Wu, K. C., Peng, D. R., Atherton, J. C. & Xue, H. (2003). *cagA* and *vacA* genotype of *Helicobacter pylori* associated with gastric diseases in Xi'an area. *World J Gastroenterol* **9**, 1762–1766.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Suerbaum, S. & Achtman, M. (2004). *Helicobacter pylori*: recombination, population structure and human migrations. *Int J Med Microbiol* **294**, 133–139.
- Suerbaum, S. & Michetti, P. (2002). *Helicobacter pylori* infection. *N Engl J Med* **347**, 1175–1186.
- Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. (1998). Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci U S A* **95**, 12619–12624.
- Suerbaum, S., Josenhans, C., Sterzenbach, T., Drescher, B., Brandt, P., Bell, M., Droge, M., Fartmann, B., Fischer, H. P. & other authors (2003). The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. *Proc Natl Acad Sci U S A* **100**, 7901–7906.
- 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. & other authors (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547.
- van de Peer, Y. & de Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* **10**, 569–570.
- van der Ende, A., Pan, Z. J., Bart, A., van der Hulst, R. W., Feller, M., Xiao, S. D., Tygat, G. N. & Dankert, J. (1998). *cagA*-positive *Helicobacter pylori* populations in China and The Netherlands are distinct. *Infect Immun* **66**, 1822–1826.
- van Doorn, L. J., Figueiredo, C., Sanna, R., Pena, S., Midolo, P., Ng, E. K., Atherton, J. C., Blaser, M. J. & Quint, W. G. (1998). Expanding allelic diversity of *Helicobacter pylori vacA*. *J Clin Microbiol* **36**, 2597–2603.
- van Doorn, L. J., Figueiredo, C., Megraud, F., Pena, S., Midolo, P., Queiroz, D. M., Carneiro, F., Vanderborght, B., Pegado, M. D. & other authors (1999). Geographic distribution of *vacA* allelic types of *Helicobacter pylori*. *Gastroenterology* **116**, 823–830.

Wang, J., van Doorn, L.-J., Robinson, P. A., Ji, X., Wang, D., Wang, Y., Ge, L., Telford, J. L. & Crabtree, J. E. (2003). Regional variation among *vacA* alleles of *Helicobacter pylori* in China. *J Clin Microbiol* **41**, 1942–1945.

Wilson, K. (1987). Preparation of genomic DNA from bacteria. In *Current Protocols in Molecular Biology*, pp. 241–245. Edited by F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. A. Smith, J. G. Seidman & K. Struh. New York: Wiley.

Wirth, T., Wang, X., Linz, B., Novick, R. P., Lum, J. K., Blaser, M., Morelli, G., Falush, D. & Achtman, M. (2004). Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proc Natl Acad Sci U S A* **101**, 4746–4751.

Yakoob, J., Fan, X. G., Peng, X. N., Hu, G. L. & Zhang, Z. (2002). *Helicobacter pylori cagA* and *vacA* cytotoxin genes in Changsha, China. *Br J Biomed Sci* **59**, 150–153.

Yamaoka, Y., Kodama, T., Kashima, K., Graham, D. Y. & Sepulveda, A. R. (1998). Variants of the 3' region of the *cagA* gene in *Helicobacter pylori* – associated diseases. *J Clin Microbiol* **36**, 2258–2263.

Zhang, Y., Argent, R. H., Letley, D. P., Thomas, R. J. & Atherton, J. C. (2005). Tyrosine phosphorylation of *cagA* from Chinese *Helicobacter pylori* isolates in AGS gastric epithelial cells. *J Clin Microbiol* **43**, 786–790.

Zhou, J., Zhang, J., Xu, C. & He, L. (2004a). *cagA* genotype and variants in Chinese *Helicobacter pylori* strains and relationship to gastroduodenal diseases. *J Med Microbiol* **53**, 231–235.

Zhou, W., Yamazaki, S., Yamakawa, A., Ohtani, M., Ito, Y., Keida, Y., Higashi, H., Hatakeyama, M., Si, J. & Azuma, T. (2004b). The diversity of *vacA* and *cagA* genes of *Helicobacter pylori* in East Asia. *FEMS Immunol Med Microbiol* **40**, 81–87.

Edited by: G. E. Duhamel