INFECTION AND IMMUNITY, Feb. 2001, p. 1160–1171 Vol. 69, No. 2
0019-9567/01/$04.00+0 DOI: 10.1128/IAI.69.2.1160–1171.2001

Allelic Variation within *Helicobacter pylori babA* and *babB*

DAVID T. PRIDE,[1]* RICHARD J. MEINERSMANN,[2] AND MARTIN J. BLASER[1]

Division of Infectious Diseases, Department of Medicine, Vanderbilt University School of Medicine and VA Medical Center, and Department of Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee,[1] and USDA Agricultural Research Service, Athens, Georgia[2]

Received 3 July 2000/Returned for modification 18 October 2000/Accepted 17 November 2000

Helicobacter pylori **strains show both geographic and disease-associated allelic variation. We investigated the diversity present in two genes, *babA* and *babB*, which are members of a paralogous family of outer membrane proteins. Eleven family members within a single *H. pylori* strain, predicted to encode proteins with substantial N- and C-terminal similarity to each other, were classified as *babA* paralogues. In their central regions, most are less than 54% related to one another. Examining the *babA* and *babB* central regions in 42 *H. pylori* strains from different geographic locales, we identified five different allele groups of *babA* (AD1 to AD5) and three different allele groups of *babB* (BD1 to BD3). Phylogenetic analysis revealed that the allelic groupings of *babA* and *babB* are independent of one another and that, for both, geographic variation is present. Analysis of synonymous and nonsynonymous substitutions in these regions showed that *babA* is more diverse, implying an earlier origin than that of the same region of *babB*, but that the *babA* diversity region may have more functional constraints. Although recombination has been central to the evolution of both genes, with *babA* and *babB* showing low mean compatibility scores and homoplasy ratios of 0.71 and 0.67, respectively, recombination is not sufficient to obscure evidence of clonal descent. Despite the involvement of *babA* in binding to the host blood group antigen Lewis B, neither the presence of different *babA* allele groups nor that of different *babB* allele groups is a determining factor in Lewis B binding of *H. pylori* strains.**

Helicobacter pylori is a microaerophilic, gram-negative bacterium that colonizes the human stomach (12). Colonization induces chronic gastritis and plays a role in the development of peptic ulcer disease and gastric adenocarcinoma (12, 44). *H. pylori* strains are highly diverse (22, 31), as evidenced by allelic variation within *vacA* (10, 11, 17, 42, 43); the presence of nonconserved DNA fragments among different strains, such as the *cag* pathogenicity island (4, 15, 16, 28, 41, 44); and the occupation of a single genomic site with different genes such as *iceA1* and *iceA2* (44). DNA fingerprinting and multilocus enzyme electrophoresis techniques have shown that there is greater total genetic diversity for *H. pylori* than for other bacteria that have been studied (2, 3, 20, 22, 32, 39).

H. pylori adhesion to the gastric epithelium is mediated, at least in part, through the Lewis B blood group antigen (13). Adherence to the epithelium is believed to help protect the bacteria from gastric acidity, as well as from displacement due to peristalsis. Two *H. pylori* genes, *babA* and *babB*, were identified based on the N-terminal similarity of their products to the Lewis B binding protein (25), and it was determined that the *babA* gene product is necessary for Lewis B binding activity. The two gene products are members of a paralogous family of outer membrane proteins, in which the members have significant N- and C-terminal similarity (40). *babA* and *babB* are nearly identical in their 5′ and 3′ regions, with most of their sequence divergence being in their midregions (8, 25).

We sought to examine the relationships between *babA* and other members of this family of proteins and to examine the diversity that exists within *babA* and *babB*. Our specific goals were to determine whether there exists geographic variation or allele group differences among *babA* and *babB* sequences and to determine whether there is any association between *babA* and *babB* diversity and Lewis B binding phenotypes of *H. pylori* strains. We also sought to examine the evolutionary relationships between *babA* and *babB* and to determine whether their current gene structures have been largely shaped by recombination.

* Corresponding author. Mailing address: Department of Medicine, New York University School of Medicine, Infectious Diseases, Veterans Affairs Medical Center, Room 6006W, 423 East 23rd St., New York, NY 10010. Phone: (212) 252-7164. Fax: (212) 252-7167. E-mail: David.T.Pride@Vanderbilt.edu.

MATERIALS AND METHODS

Strains and growth conditions. A total of 42 *H. pylori* strains were obtained from patients from several different locales and with differing clinical diagnoses (Table 1). Strains were cultured at 37°C in a 5% CO_2 atmosphere on Trypticase soy agar plates with 5% sheep blood (TSAB) and frozen at −70°C until used in this study.

Genotyping of *H. pylori*. Genomic DNA was prepared from each strain after 48 to 72 h of growth on TSAB plates, as described previously (47). Strains were examined by PCR to determine *vacA* allelic types and the presence of *cagA, iceA1*, and *iceA2*, using established primers (10, 43, 48). PCR products were analyzed by agarose gel electrophoresis (10).

Analysis of *babA* homologs. To compare sequences of *babA* and other homologs from strains J99 and 26695, nucleotide searches were performed using BLAST algorithms (9) from GenBank and The Institute for Genomic Research (http://www.tigr.org/tdb/mdb/hpdb/hpdb.html) and Astra-Boston (http://sciabin.astrazeneca-boston.com/hpylori/) databases. Translations of nucleotide sequences were performed using GCG (Genetics Computer Group; Madison, Wis.) Translate, and pairwise similarities were determined using GCG Gap. Points of division were determined based on an alignment of the paralogues, where the first division point corresponds to the beginning of the HP722 open reading frame (ORF) (nucleotides 1 to 360 of 26695 *babA*), the second division point corresponds to the beginning of the HP229 ORF (nucleotides 361 to 760), and the third division point is based on the midpoint of the HP229 ORF (nucleotides 761 to 1510 and nucleotide 1511 to the end of each ORF). Secondary structure analysis was performed using Garnier's algorithm (21) at http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor.html, transmembrane domains were deduced using the TMPred server at http://www.ch.embnet

TABLE 1. *H. pylori* strains used in this study

Strain[a]	*vacA* type		*iceA* type	Allele group		Clinical diagnosis[c]
	s	m		*babA*	*babB*	
J99	s1b	m1	2	AD2	BD2	DU
96-74	s1b	m1	1	AD2	BD2	DU
97-141	s1b	m1	1	AD2	BD2	Gastritis
84-183	s1b	m1	2	AD2	BD3	NA
96-71	s1a	m1	2	AD1	BD2	DU
95-76[b]	s2	m2	2	AD1	BD2	Gastritis
98-292	s1b	m2	1	AD1	BD3	NA
97-503	s2	m2	2	AD1	BD3	Gastritis
95-56[b]	s2	m2	2	AD2	BD2	Gastritis
96-68	s1a	m1	1	AD3	BD2	Gastritis
98-454	s2	m2	2	AD3	BD2	NA
97-5	s1b	m2	1	AD3	BD3	Gastritis
90-40	s1a	m1	1	AD2	BD1	NA
97-147[b]	s2	m2	1	AD5	BD2	NA
97-780	s2	m2	2	AD3	BD2	NA
95-97	s1b	m1	2	AD3	BD2	Gastritis
26695	s1a	m1	1	AD2	BD3	Gastritis
96-28	s1b	m2	2	AD3	BD2	DU
CCUG17875	s1a	m1	1	AD3	BD3	NA
95-23	s1a	m1	1	AD3	BD2	NA
95-10	s1a	m1	1	AD3	BD1	NA
88-28	s1a	m2	2	AD3	BD3	NA
98-10	s1a	m1	1	AD2	BD1	GU
98-16	s1a	m1	1	AD2	BD2	GU
98-24	s1a	m1	2	AD2	BD1	DU, GU
98-18	s1a	m1	1	AD2	BD3	GU
96-10	s1a	m1	1	AD3	BD2	DU
98-22	s1a	m1	2	AD3	BD3	Gastritis
98-26	s1a	m1	1	AD3	BD2	GU
98-32	s1a	m1	2	AD3	BD3	DU, GU
98-30	s1a	m1	1	AD3	BD2	GU
98-14	s1a	m1	1	AD2	BD1	DU
98-12	s1a	m1	1	AD4	BD1	DU
97-793	s1a	m1	1	AD2	BD2	Gastritis
97-776	s1a	m2	1	AD2	BD2	Gastritis
98-60	s1a	m1	1	AD4	BD2	Gastritis
98-344	s1c	m2	2	AD3	BD2	Gastritis
98-77	s1a	m1	1	AD3	BD2	Gastritis
97-679	s1a	m1	2	AD4	BD2	Gastritis
97-723	s1a	m2	2	AD4	BD3	Gastritis
97-687	s1a	m1	2	AD3	BD2	Gastritis
98-53	s1a	m2	2	AD3	BD2	Gastritis

[a] The geographic location of strains is outlined in Fig. 4.
[b] *cagA*-negative strain.
[c] NA, not available; DU, duodenal ulcer; GU, gastric ulcer.

.org/software/TMPRED_form.html (24), and isoelectric points were deduced using the server at http://www.expasy.ch/tools/pi_tool.html (46).

Diversity region fragment amplification and sequence analysis. *babB* fragments were amplified using primers AN5036 and AN5037 (Table 2). *babA* fragments were amplified primarily with primers C6678 and C6679; however, alternate combinations of forward and reverse primers among C6678, C6679, B1999, BA909, and B1998 were used to amplify fragments from strains that did not amplify with the primary primer set. PCR fragments were purified using the QiaQuick PCR purification kit and the QiaQuick Gel Extraction kit. PCR fragments were directly sequenced on both strands, using an automated Applied Biosystems, Inc., sequencer at the Vanderbilt Cancer Center sequencing core facility, and were analyzed using Sequencer 3.1.1 (Gene Codes Corp, Inc., Ann Arbor, Mich.).

Phylogenetic and nucleotide analyses. Multiple alignments of nucleotide and predicted amino acid sequences were created using GCG Pileup, ClustalW, and GCG Pretty (Wisconsin Package version 9.1; Genetics Computer Group). Similarity plots were created using SimPlot 2.5 (http://www.welch.jhu.edu/šray/download). Phylograms of each nucleotide alignment were generated using both parsimony and distance matrix methods, using Paup 4.0b2 (Sinauer Associates, Sunderland, Mass.). All phylograms were displayed using Treeview (35) and Paup 3.1 (Illinois Natural History Survey, Champaign), using midpoint rooting. Transitions and transversions were calculated with MEGA 1.01 (Pennsylvania State University, University Park). Nucleotide and amino acid similarities of both *babA* and *babB* fragments were determined using BoxShade 3.2. Average GC content was calculated using GeneDoc 2.5 (K. B. Nicholas and H. B. Nicholas, Jr.). Synonymous and nonsynonymous substitution rates were generated using DnaSP 3.0 (37). The amount of homoplasy (*H* ratio) was determined as the mean of five independent tests using the homoplasy test (33), and all sites containing gaps were excluded from the analysis. The consistency index was determined using MacClade Version 3.0 (Sinauer Associates). Compatibility matrices were created using Reticulate (26). All means, standard deviations, and ratios were determined using Corel Quatro Pro (Corel Corp., Ottawa, Canada).

***babA* and *babB* mutant strains.** The *babA* and *babB* ORFs were amplified from *H. pylori* strain J99, using primer pairs B9617-AN5954 and C5773-C5774, respectively (Table 2). Each ORF was cloned into pGem-T Easy and transformed into *Escherichia coli* DH5α. A unique *Xba*I site was used to linearize the resulting plasmids containing *babA*, and a unique *Bam*HI site was used to linearize *babB*-containing plasmids. The *aphA* cassette (kanamycin resistance) was amplified from plasmid pUC4K using primer pair C7491-C7492 or AN5485-AN5486 with either *Xba*I or *Bam*HI restriction ends, respectively (Table 2). The resulting fragments were then cloned into the respective unique sites in *babA* and *babB*, creating plasmids pDP501 and pDP601. Strain J99 was transformed to Kan with both plasmids, and chromosomal DNA was isolated from each resulting strain. Insertion of the *aphA* cassette into *babA* or *babB* was confirmed using primer pairs AN5045-AN5954 and AN5043-C5774, respectively.

Lewis B binding assay. *H. pylori* strains were cultured for 48 h at 37°C in a 5% CO_2 atmosphere on TSAB plates. Cells were harvested into 50 ml of brucella broth supplemented with 5% newborn calf serum and grown for another 48 h. Cells then were harvested and diluted to an optical density (OD) of 1.0 at 450 nm. One milliliter of each culture was then resuspended in a Phosphate-buffered saline (PBS) solution containing Thimersol and Tween 20 (PBSTT), plus 0.1% gelatin, and incubated at 37°C for 1 h. Cells were washed in PBSTT and incubated with or without 1 μg of Lewis B-human serum albumin glycoconjugate (Isosep, Tullinge, Sweden) for 1 h at 37°C in PBS containing 0.5% bovine serum albumin. Cells were washed, followed by incubation for 1 h at 37°C with anti-Lewis B murine monoclonal antibody (Signet Pathology Laboratories, Inc., Dedham, Mass.) in PBS containing 0.1% gelatin and 0.5% bovine gamma globulin. Cells were washed, followed by incubation with anti-mouse immunoglobulin M horseradish peroxidase conjugate (Cappel Laboratories, Cochranville, Pa.) for 1 h at 37°C in a solution containing 0.1% bovine gamma globulin and 1% bovine serum albumin. Cells were washed twice and developed for 30 min in 500 μl of a solution containing 9.4 ml of 0.2 M Na_2HPO_4, 10.6 ml of 0.1 M citric acid, and 32 μl of H_2O_2. Results were expressed in relative units (OD × 1,000) as the OD difference between cells with Lewis B and those incubated without Lewis B. Each experiment was repeated on consecutive days.

Nucleotide sequence accession numbers. *babA* GenBank accession numbers are AF277904 to AF277942, and *babB* GenBank accession numbers are AF277943 to AF277981.

RESULTS

Classification of *babA* paralogues. A GenBank BLAST search of the genome in *H. pylori* strain 26695 using *babA* (HP1243) nucleotides revealed 24 genes with scores of >100 and probabilities ranging from 0.0 to 0.998. For 11 genes, probabilities ranged from 0.0×10^{-118} to 1.9×10^{-118}, whereas for the other 13 genes, probabilities were substantially lower. Using *babA* (JHP833) from *H. pylori* strain J99 to search that strain's ORFs, 22 genes were identified with scores of >100, with 10 having probabilities ranging from 0.0×10^{-95} to 7.5×10^{-95} and the other 12 genes having substantially lower scores. Therefore, our further analysis concentrated on the strongest paralogues in each strain.

All of the 11 genes from 26695 are members of a paralogous family, previously identified as outer membrane proteins through the characteristic C-terminal alternating hydrophobic residues (40), with at least one N-terminal domain of similarity and at least seven C-terminal domains of similarity. Two genes (HP25 and HP229) are predicted on the basis of sequence homology to encode porins (18, 19). These 11 genes

TABLE 2. PCR primers used for this study

Primer name	Gene or designation	Direction	Location in gene[c]	GenBank accession no.	Sequence (5′-3′)[b]
C6778	*babA* (HP1243)	Forward	570–588	AE000629	GCTTACCCGCGCTCAAAG
C6779	*babA* (HP1243)	Reverse	1056–1076	AE000629	CTCCGTGAAAGGGTTGAAAG
B1999	*babA* (HP1243)	Reverse	1088–1107	AE000629	GTTAAGCGAGCATGCCTTG
BA909	*babA* (HP1243)	Forward	410–428	AE000629	CTTCAACCACCATCTTCA
B1998	*babA* (95-76)[a]	Forward	562–580	NA[a]	GCTTGCCAGCGCTCAACC
AN5036	*babB* (HP896)	Forward	451–469	AE000599	ACCATCACTTGCAATTCG
AN5037	*babB* (HP896)	Reverse	1042–1060	AE000599	GAGCGTTTTTGAGCATGC
B9617	*babA* (HP1243)	Forward	1–23	AE000629	ATGAAAAAACACATCCTTTCAT
AN5954	*babA* (JHP833)	Reverse	2213–2235	AE001512	TTAGTAAGCGAACACGTAATTC
C5773	*babB* (HP896)	Forward	1–20	AE000599	ATGAAAAAAACCCTTTTAC
C5774	*babB* (HP896)	Reverse	2138–2156	AE000599	TTAGTAAGCGAACACATA
AN5045	JHP834	Forward	1–19	AE001512	GTGTGCGGCGTATTATCG
AN5043	HP1244	Forward	1–21	AE000629	ATGGAAAGAAAACGCTATTC
C7491	pUC4K	Forward	369–401	X06404	GCTCTAGAGCTCACGACGTTGTAAAACGACGGCCAGTG
C7492	pUC4K	Reverse	1598–1638	X06404	GCTCTAGAGCGTTGTGTCTCAAAATCTCTGATGTTAC
AN5485	pUC4K	Forward	369–401	X06404	CGCGGATCCGCGTCACGACGTTGTAAAACGACGGCCAGTG
AN5486	pUC4K	Reverse	1598–1638	X06404	CGCGGATCCGCGGTTGTGTCTCAAAATCTCTGATGTTAC

[a] NA, not available, unpublished data.
[b] Underlined sequences correspond to *Xba*I and *Bam*HI restriction sites.
[c] Nucleotide positions.

(10 genes from J99) had significant 5′ similarity (over 350 nucleotides) but much greater 3′ similarity (over 750 to 1,500 nucleotides).

These 11 genes (10 genes in J99) share other properties. By Garnier's secondary structure analysis, each protein sequence was predicted to have parallel motifs at similar relative positions (data not shown). All have predicted basic pIs ranging from 8.47 to 9.24 (data not shown), except for two paralogues encoded by JHP212 and JHP1261 from strain J99 (both with pIs of 6.52). Each ORF has at least two predicted transmembrane domains (one each within the 3′ and 5′ regions), and all contain a predicted N-terminal signal peptidase I signal sequence (40). Based on the BLAST search probability scores and the similarity of the gene motifs and specific structures, we have classified these 11 genes (10 genes in J99) as *babA* paralogues.

Analysis of sequence similarities among the 26695 and J99 *babA* paralogues. Comparison of the 26695 *babA* paralogues with reference gene *babA* (HP1243) from 26695 (data not shown) showed that the first segment (nucleotides 1 to 360) is well conserved among five genes; however, the other five genes either have a truncated version or do not possess this segment. Pairwise similarity comparisons show that HP227 and HP1342 are identical throughout their ORFs and that HP317 is nearly identical to *babA* (HP1243) in this first segment; most of the paralogues have 61 to 100% similarity to each other in this same region (Table 3). The next segment (nucleotides 361 to 760) is more variable in that none are >64% similar to one another, except for HP725 and HP722 (73% identical to each other), and the third segment (nucleotides 761 to 1510) also has substantial variability (Table 4). However, the fourth segment (nucleotide 1511 to the end of each ORF) showed similarities ranging from 67 to 100% among the different paralogues. These analyses indicate that there is substantial 5′ and 3′ conservation of these paralogues, with the greatest variation in the midregions. The mean (± standard deviation) (both calculated with Corel Quatro Pro) percent nucleotide similarities for the various regions of the *babA* paralogues were as follows: nucleotides 1 to 360, (68.0 ± 14.5)%; nucleotides 361 to 760, (42.9 ± 13.2)%; nucleotides 761 to 1510, (48.5 ± 13.7)%; and nucleotide 1511 to the end, (76.0 ± 8.11)%. A

TABLE 3. Pairwise nucleotide similarity matrix of the 26695 *babA* paralogues for the 5′ regions

Paralogue (nucleotides 361 to 760)	% Similarity for paralogue (nucleotides 1 to 360)[a]										
	HP1243	HP317	HP896	HP25	HP9	HP227	HP229	HP722	HP725	HP1177	HP1342
HP1243		**98**	77	76	76	66	—[b]	—	74	43	66
HP317	48		76	76	76	65	—	—	74	43	65
HP896	43	49		72	74	70	—	—	79	73	70
HP25	40	46	64		68	63	—	—	83	48	63
HP9	35	36	37	41		74	—	—	41	44	74
HP227	47	47	54	51	40		—	—	74	61	**100**
HP229	—	—	—	—	—	—		—	—	—	—
HP722	40	36	36	39	36	36	—		—	—	—
HP725	36	37	37	41	44	38	—	73		32	74
HP1177	42	40	34	36	49	38	—	44	41		61
HP1342	47	47	54	51	40	**100**	—	36	38	38	

[a] Percent similarity was determined by using the GCG gap program. Each value was rounded to the nearest integer, and values of >85% are represented in boldface.
[b] —, no analysis was possible, since there was no specified sequence due to alignment of paralogues.

TABLE 4. Pairwise nucleotide similarity matrix of the 26695 *babA* paralogues for the 3′ regions

Paralogue (nucleotide 7511 to end)	% Similarity for paralogue (nucleotides 761 to 1510)[a]										
	HP1243	HP317	HP896	HP25	HP9	HP227	HP229	HP722	HP725	HP1177	HP1342
HP1243		62	72	60	56	52	40	38	38	38	52
HP317	**100**		75	61	63	48	40	37	38	39	48
HP896	**97**	**97**		73	58	53	39	44	44	39	53
HP25	72	72	73		58	53	35	43	42	39	53
HP9	78	77	80	75		56	38	41	41	54	56
HP227	78	78	79	72	79		38	40	38	38	**100**
HP229	69	67	72	72	71	71		42	38	40	38
HP722	69	69	74	74	69	76	68		**89**	38	40
HP725	69	69	74	74	70	76	67	**100**		41	38
HP1177	71	72	76	74	73	79	70	74	74		38
HP1342	78	78	79	72	79	**100**	71	76	76	79	

[a] Percent similarity was determined using the GCG gap program. Each value was rounded to the nearest integer, and values of >85% are represented in boldface.

parallel analysis of the 10 *babA* paralogues in strain J99 identified similar relationships (data not shown).

Paralogue-flanking genes and intergenic regions. Most of the *babA* paralogues are in similar genomic locations in 26695 and J99, based on their flanking genes. Each is flanked by at least one gene that is identical between the strains with the exception of *babA, babB*, and HP25 (data not shown). Although the genes flanking *babA* in 26695 flank *babB* in J99, the genes that flank *babB* in 26695 do not immediately flank *babA* in J99. The only significant identity in the upstream intergenic regions exists between *babA* (HP1243) and HP317 (74% over 300 nucleotides), its most similar paralogue in strain 26695, and between HP722 and HP725 (88% over 290 nucleotides). In contrast, there is substantial (74 to 98%) identity in the intergenic regions downstream of *babA, babB*, and HP317. Although *babA* and HP317 are most similar, the regions downstream of *babB* and HP317 are 98% identical. Since HP317 and *babB* are flanked by identical downstream genes (HP316 and HP895), there is virtual identity for 900 bp downstream (data not shown).

Phylogenetic analysis of *babA* paralogues. To better understand the origins of these genes, the phylogeny of the 21 paralogues from the two strains was determined (Fig. 1). Each 26695 ORF is most closely related to an ORF in J99, indicating the substantial interstrain conservation of each paralogue. The identical ORFs in 26695 (HP227 and HP1342) are closely related to the identical ORFs in J99 (JHP212 and JHP1261). ORF HP317 has no J99 counterpart. The phylogeny confirms the expectation that *babA*, HP317, and *babB* all are highly related and that HP722 and HP725 are highly related to each other. All bootstrap values on the phylogram are ≥75, and essentially identical phylogenies were produced using maximum parsimony algorithms (data not shown).

Although the genomes of 26695 and J99 have an overall 94% conservation at the nucleotide level (5), the 10 paired *babA* paralogues ([90.4 ± 2.4]%) are less highly related (Fig. 1). Levels of amino acid identity generally parallel the nucleotide identity for each pair of homologs (data not shown). The average GC content for the *H. pylori* genome is 39% (5, 40), while for the 11 paralogues in 26695, the mean GC content is (42.9 ± 1.6)% (range from 41 to 46%). For the J99 paralogues, the mean GC content is (43.1 ± 1.6)% (range from 41 to 45%). The GC content of the cluster (Fig. 1, cluster A) of paralogues including *hopA* is (41.4 ± 0.6)%, whereas for the cluster (Fig. 1, cluster B) of paralogues including *babA* it is (44.2 ± 1.0)%. This significant ($P < 0.001$, Student's t test) difference is further support for the validity of the phylogram.

Allelic diversity within *babA*. Because of the segmental conservation and diversity among these paralogues, we next examined the best-characterized paralogues, *babA* and *babB*, for their diversity among different *H. pylori* strains. From multiple alignments of the four and three sequences of *babA* and *babB* available from different strains, respectively (5, 25, 40; unpublished data), there is strong 5′ and 3′ conservation for both genes, with the greatest variation occurring in the midregions (data not shown). The greatest diversity occurs between nucleotides 612 and 1046 in *babA* (86% mean identity) and between 488 and 1010 in *babB* (91% mean identity). We then examined these *babA* and *babB* regions by directly sequencing PCR products from 42 *H. pylori* strains from around the world (Table 1). The *babA* sequences from the different strains are highly variable (Fig. 2); however, the segment including predicted amino acids 306 to 334 (nucleotides 915 to 999) shows five distinct families of variants, designated allele groups AD1 (*babA* diversity allele 1), AD2, AD3, AD4, and AD5 (Table 1), as determined through similarity plot analysis (Fig. 3A and data not shown). However, there exists a spectrum of similarity among the different *babA* allele groups, where groups AD3 and AD4 are most similar to one another, followed by group AD2, while groups AD1 and AD5 have little overall similarity to any of the other allele groups (Fig. 3B and data not shown). Phylogenetic analyses of nucleotides 915 to 999 also support the grouping of five separate *babA* variant families (Fig. 4, segment 1 phylogram), yielding results congruent with those of the similarity plots, while parallel analyses of the flanking nucleotides (830 to 914 and 1000 to 1084) do not show these allelic groupings (Fig. 4 and data not shown). All 42 strains belong to one of these allele groups, with AD5 strain 97-147 resembling each of the other four allele groups (Fig. 2).

Allelic variation within *babB*. Analysis of the translated *babB* sequences (nucleotides 488 to 1010 of 26695 *babB*) suggests that *babB* is more conserved than is *babA*; however, the region between amino acids 216 and 252 (nucleotides 646 to 754) shows three families of variants, designated allele groups BD1 (*babB* diversity allele 1), BD2, and BD3 (Fig. 5 and Table 1), as initially discovered through similarity plot analysis (Fig. 3C and data not shown). As for *babA*, there also exists a spectrum of similarity, where allele groups BD2 and BD3 are

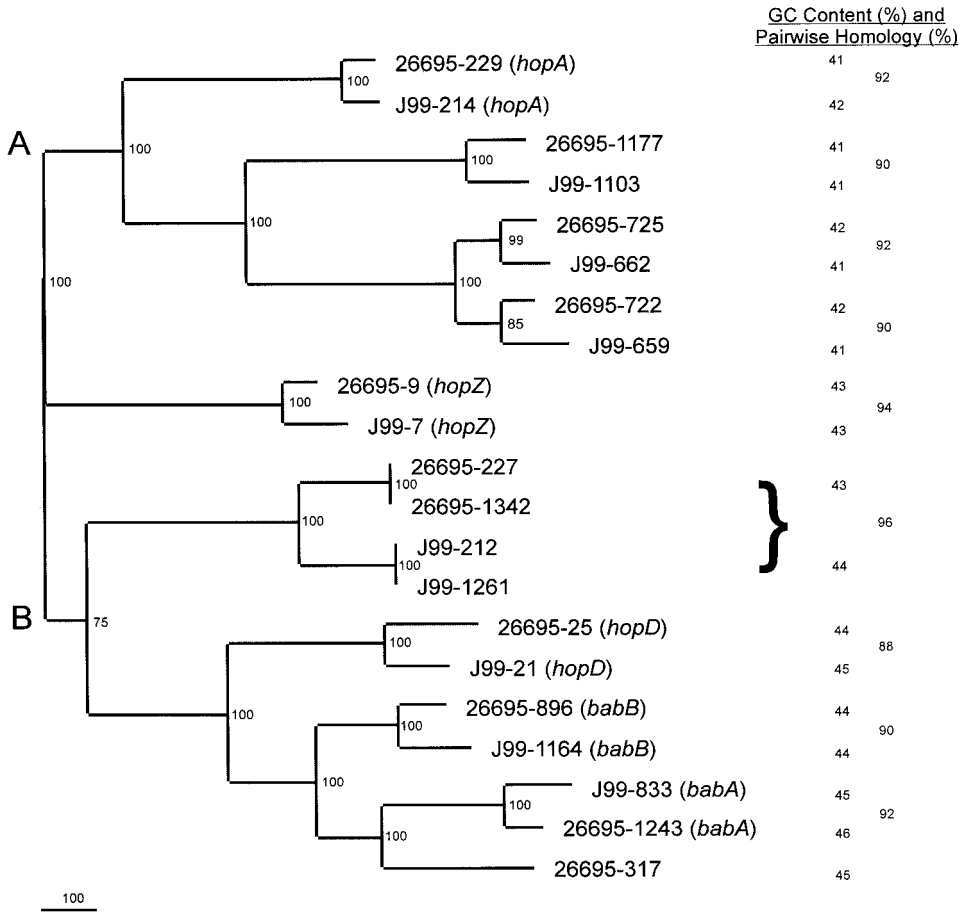

FIG. 1. Nucleotide phylogram of *babA* paralogues from *H. pylori* strains 26695 and J99. Sequences were obtained from The Institute for Genomic Research database (26695) and the database at Astra-Boston (J99), aligned using GCG Pileup, and analyzed using PAUP 4.0b2 neighbor-joining analysis based on Kimura's two-parameter model distance matrices. Bootstrap values (based on 500 replicates) are represented at each node, and the branch length index is represented below the tree. Strain names and gene numbers are indicated at the termination of each branch. The percent nucleotide similarity (determined using GCG Gap) of each pair of homologs is indicated in the right column. The brace is used to indicate the similarity between the two sets of paired ORFs from each strain. The GC content of each ORF is shown in the middle column. The mean GC content of the paralogues of the top (A) cluster of (41.4 ± 0.6)% is significantly lower than that for the bottom (B) cluster ([44.2 ± 1.0]%; $P < 0.001$).

most similar to one another, while group BD1 has little overall similarity to the other allele groups (Fig. 3D). Phylogenetic analysis also reinforces these groupings (Fig. 6, segment 1 phylogram), but these groupings are not present in the flanking (537 to 645 and 755 to 863) nucleotides. Importantly, strains that are clustered based on particular *babA* allelic types are not clustered based on their *babB* allelic types (Fig. 4 and 6).

Geographic variation within *babA* and *babB*. Phylogenetic analysis of the regions 3′ (segment 3) of the *babA* and *babB* allele group regions demonstrates geographic clustering of the strains (Fig. 4 and 6). All but three of the U.S. strains cluster on the upper half of the *babA* segment 3 phylogram, while two of the remaining three U.S. strains cluster near three of the four South American strains on the lower half. Most strains from Europe and Oceania cluster with the U.S. strains. Many of the Indian strains form a small cluster on the upper half, and all of the East Asian strains are found on the lower half of the phylogram. In the phylogram of *babB* segment 3, all of the Japanese strains are present on the upper branch, whereas all but one of the U.S. and South American strains are on the lower half. Thus, for both *babA* and *babB*, allele groupings (segment 1) that are not related by geographic origin of the strains and that are 5′ of regions (segment 3) that are largely related according to geographic origin are present. The regions 5′ of segment 1 for both genes also show evidence of geographic clustering (data not shown).

Comparison of *babA* and *babB* diversity regions. Among the 42 strains studied, the *babA* diversity region varied in length from 434 to 488 bp (>10%), whereas there was little *babB* length heterogeneity (514 to 526 bp). The *babB* diversity region also was much more conserved at both the nucleotide (87 versus 78%) and amino acid (84 versus 74%) levels than was the *babA* diversity region. Nearly every *babA* and *babB* fragment was unique, except that Indian strains 98-344 and 97-679 had identical *babB* fragments while differing in other genotypes (Table 1). The GC contents for the fragments (46 and 43% for *babA* and *babB*, respectively) resembled those for the entire genes (Fig. 1). Both transitions and transversions occurred at high levels but more often in *babA* than in *babB* (data not shown), further indicating the diversity in these regions.

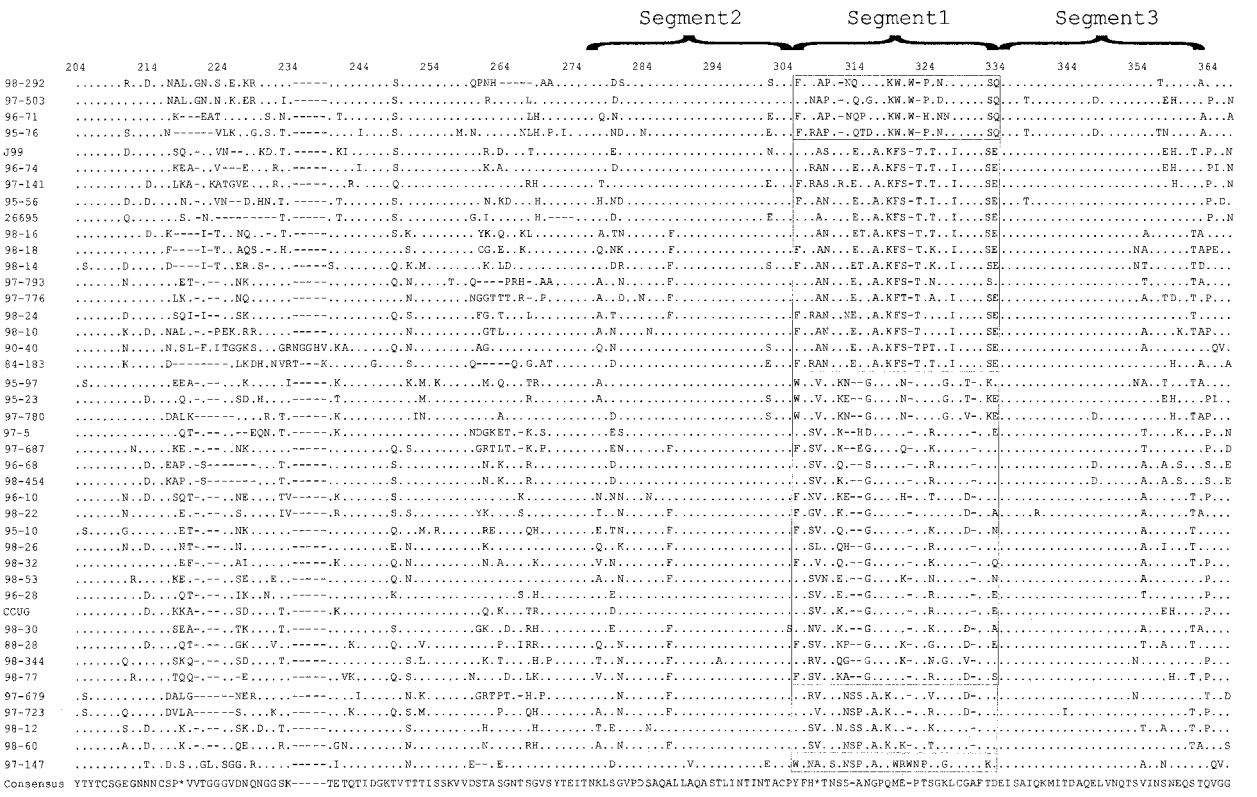

FIG. 2. Amino acid alignment of 42 *babA* diversity region fragments. Amino acid translations of nucleotide sequences were performed using GCG Translate, and sequences were aligned using GCG Pileup. A consensus sequence (based on a plurality of 10) was determined using GCG Pretty, and all the sequences were displayed using Boxshade 3.2. The dots refer to sites where amino acids match those of the consensus sequence, the hyphens represent deletions, the asterisks represent sites where the consensus sequence is indeterminate, and the boxes are used to separate different allele groups among the *babA* sequences. Each allele group is based on amino acids 306 to 334 and is represented at the right of the alignment. The *babA* diversity alleles 1 to 5 are represented as AD1 to AD5, respectively.

Analysis of synonymous and nonsynonymous substitutions in the *babA* and *babB* diversity regions. To explore the evolutionary basis for variation in the *babA* and *babB* diversity regions, we analyzed K_s (mean number of synonymous substitutions per site), K_a (mean number of non synonymous substitutions per site), and the K_a/K_s ratio for all possible pairs of aligned sequences. Since synonymous base changes do not alter amino acid translations, their usage is less constrained than that of nonsynonymous base changes, and thus K_s should roughly reflect the divergence time between sequences. The greater value of K_s for *babA* than for *babB* (0.33 versus 0.23) suggests a more distant common ancestor for *babA*. Since the K_a/K_s ratio controls for differences in divergence time, the ratios' similarity for *babA* and *babB* (0.42 versus 0.51) indicates that these two regions are subject to parallel functional constraints, which are less than previously determined within single alleles for the 5′ region of *cagA* (42) and the midregion of *vacA* (11). Comparisons of K_a/K_s ratios among the *babA* or *babB* allele groups indicate that the functional constraints for each allele group are similar (data not shown).

Homoplasy and recombination in the *babA* and *babB* fragments. Homoplasy is similarity within a locus that is not attributable to common ancestry and may reflect convergent, parallel, or reverse evolution (MacClade). In the absence of specific selective influences, recombination is believed to be the most important mechanism resulting in homoplasies on bacteria. The consistency index measures the amount of homoplasies in a population (from 0.0 to 1.0), where 1.0 indicates the absence of homoplasy. Phylograms of the *babA* and *babB* diversity regions have identical consistency indices of 0.34, indicating that there is substantial homoplasy among both these sequences. To determine whether the homoplasy is due to recombination, both the *babA* and *babB* diversity region fragments then were examined by the homoplasy test (33); which measures the frequency with which the same nucleotide changes (homoplasies) occur in different branches of a maximum parsimony tree. The product, the *H* ratio, represents the frequency of observed synonymous homoplasies relative to the frequencies expected for the observed sequence variation under clonal descent (in the absence of recombination, $H = 0$) or free recombination ($H = 1$) (1, 33, 38). Several *H. pylori* housekeeping genes yield ratios ranging from 0.60 to 0.79, indicating high levels of recombination for these genes (1). Although *vacA* gene fragments yield an *H* ratio of 0.69, *cagA* gene fragments yield *H* ratios of 0.15 to 0.17, among the lowest ratios yet observed for any species (38). *babA* and *babB* fragments yielded *H* ratios of 0.71 and 0.67, respectively, indicating that recombination is frequent in both of these loci and of a magnitude similar to that in the housekeeping genes.

As a further measure of the extent of recombination within *babA* and *babB*, compatibility matrices were used. In compat-

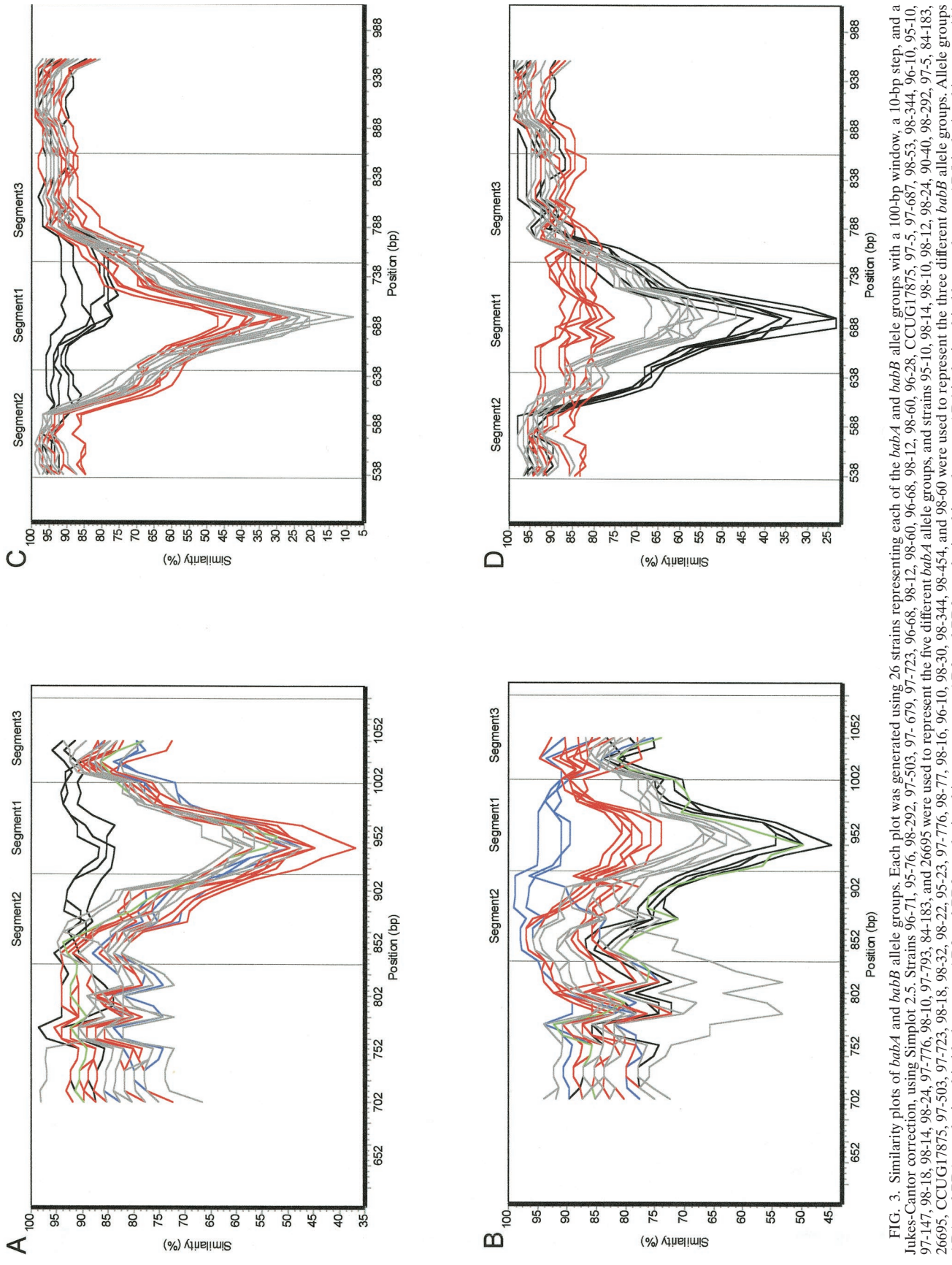

FIG. 3. Similarity plots of *babA* and *babB* allele groups. Each plot was generated using 26 strains representing each of the *babA* and *babB* allele groups with a 100-bp window, a 10-bp step, and a Jukes-Cantor correction, using Simplot 2.5. Strains 96-71, 95-76, 98-292, 97-503, 97- 679, 97-723, 96-68, 98-12, 98-60, 96-68, 98-12, 98-60, 96-28, CCUG17875, 97-5, 97-687, 98-53, 98-344, 96-10, 95-10, 97-147, 98-18, 98-14, 98-24, 97-776, 98-10, 97-793, 84-183, and 26695 were used to represent the five different *babA* allele groups, and strains 95-10, 98-14, 98-10, 98-12, 98-24, 90-40, 98-292, 97-5, 84-183, 26695, CCUG17875, 97-503, 97-723, 98-18, 98-32, 98-22, 95-23, 97-776, 98-77, 98-16, 96-10, 98-30, 98-344, 98-454, and 98-60 were used to represent the three different *babB* allele groups. Allele groups AD1 and BD1 are depicted in black, groups AD2 and BD2 are depicted in gray, groups AD3 and BD3 are depicted in red, group AD4 is depicted in blue, and group AD5 is depicted in green. (A) Plot of *babA* sequences, using 96-71 (AD1) as the reference strain. (B) Plot of *babA* sequences, using 97-679 (AD4) as the reference strain. (C) Plot of *babB* sequences, using 95-10 (BD1) as the reference strain. (D) Plot of *babB* sequences, using 98-18 (BD3) as the reference strain.

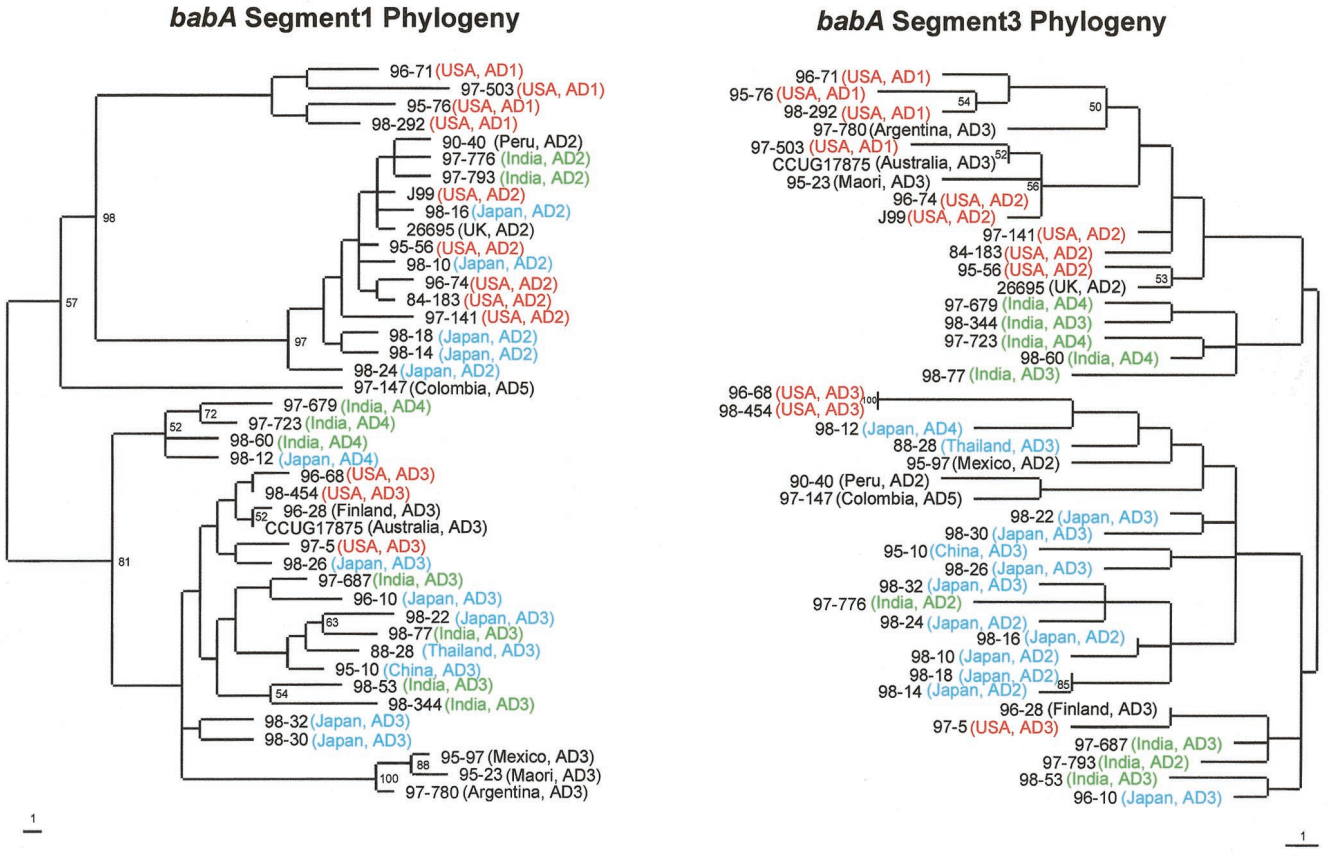

FIG. 4. Phylograms of *babA* segment 1 (nucleotides 915 to 999 on the left) and *babA* segment 3 (nucleotides 1000 to 1084 on the right). Segment 1 includes the *babA* allele group-defining sequences, and segment 3 is the region downstream of segment 1. Regions from 42 strains underwent sequence analysis as previously described and were aligned using GCG Pileup, and specific regions were identified using GCG Pretty. The sequences then were realigned using ClustalW and subjected to phylogenetic analysis using Paup 4.0b2 neighbor-joining analysis based on Kimura's two-parameter model distance matrices. Strain number, geographic location, and allelic type (AD1 to AD5) are represented at the termination of each branch, and the branch length index is shown below each tree. Bootstrap values of >50 (based on 500 replicates) are shown at each node. Colors are used to represent strains from the United States (orange), Asia (other than India) (blue), India (green), and other countries or groups (black).

ibility matrices, white spaces represent pairs of informative sites that are compatible with a maximal parsimony tree, whereas black spaces show pairs of sites that are incompatible. The mean compatibility score measures the average frequency with which pairs of informative sites are compatible. Both the *babA* and *babB* matrices are largely filled with black spaces (Fig. 7), and both yield similarly low mean compatibility scores (0.31 and 0.34, respectively), confirming that repeated recombination events have been involved in the evolution of both gene segments.

Lewis B binding for *babA* and *babB* allele groups. To determine if there is any association between *babA* and *babB* allele groups and the ability of the strains to bind Lewis B, the Lewis B binding capacity of a group of strains representing each allele group was assessed. There was substantial variation from <10 to 658 units (mean, 346 ± 282) (see Materials and Methods for description of units) among these strains; however, there was no association between the allele group status of strains from each allele group and Lewis B binding (data not shown). Each *babB* allele group included strains that bound Lewis B well, or not at all, while strains from each *babA* allele group showed a broad range of values. Isogenic *babA* and *babB* mutants also were tested in the assay (Table 5). The *babA* mutants failed to bind Lewis B, whereas the *babB* mutants bound to a similar degree as did the wild-type strains (Table 5). These results further indicate that, while *babA* is involved in binding to Lewis B (25), *babB* is not and that the particular *babA* and *babB* allele groups are not determining factors in Lewis B binding.

DISCUSSION

We classified the 11 (10 in J99) *babA*-related genes in 26695 as *babA* paralogues by several criteria, including the natural break found in the BLAST search probability scores in both strains J99 and 26695 and their substantial N- and C-terminal similarity. Both *babA* (HP1243) and *hopZ* (HP9) have been shown to be involved in *H. pylori* adherence to gastric cells (25, 36), and paralogues *hopD* (HP25) and *hopA* (HP229) appear to be porins, involved in molecular transport across the *H. pylori* membrane (40). These 11 paralogues do not constitute the full *H. pylori* repertoire for adherence or molecular transport, since other members (HP912 and HP913) of the family of outer membrane proteins have been shown previously to be involved in adherence (34). HP912, HP913, and HP706 also appear to function as porins (19, 40).

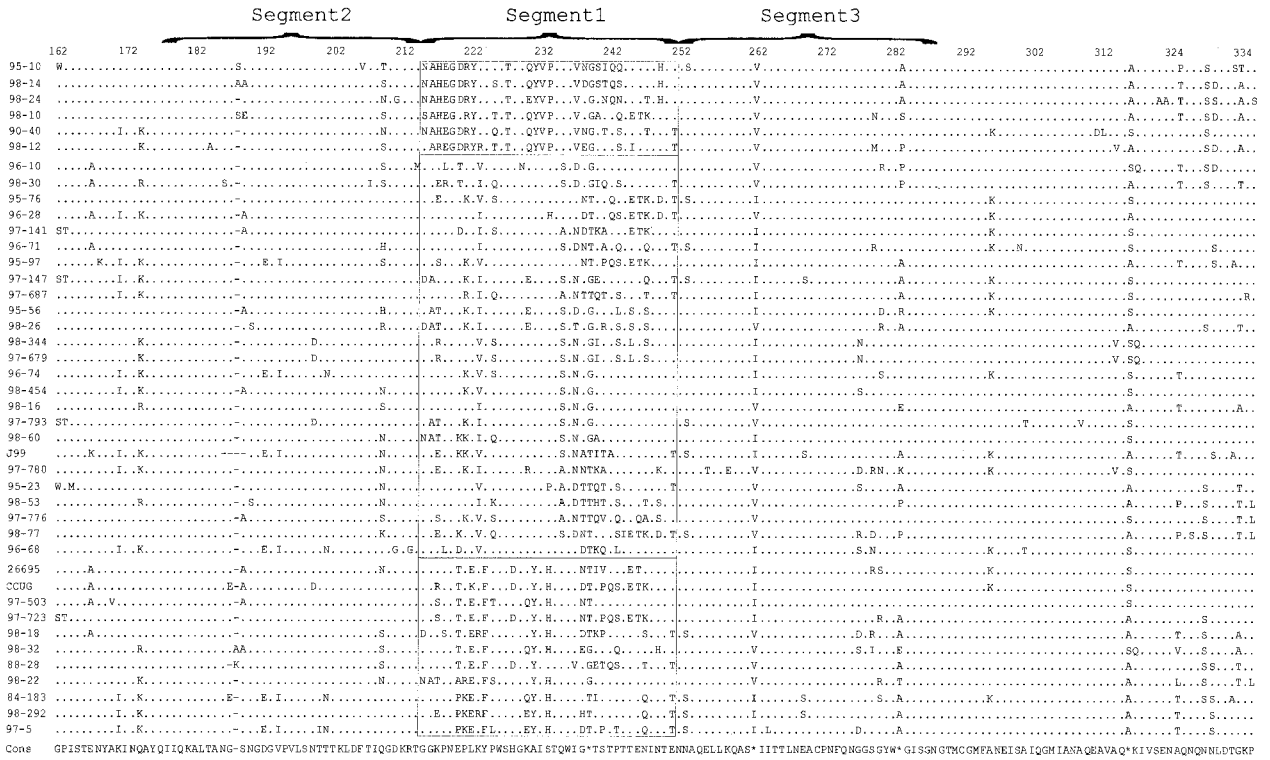

FIG. 5. Amino acid alignment of *babB* diversity regions from 42 different *H. pylori* strains. Amino acid translations of nucleotide sequences were performed using GCG Translate, and sequences were aligned using GCG Pileup. A consensus sequence (based on a plurality of 10) was determined using GCG Pretty, and all the sequences then were displayed using Boxshade 3.2. The dots refer to sites where amino acids match those of the consensus sequence, the hyphens represent deletions, the asterisks represent sites where the consensus sequence is indeterminate, and the boxes are used to separate different allele groups among the *babB* sequences. Each allele group is based on amino acids 216 to 252 and is represented at the right of each alignment. The *babB* diversity alleles 1 to 3 are represented as BD1 to BD3, respectively.

The strong similarities at both the N and C termini of most of these paralogues imply their necessity for conserved functions, whereas the central variable regions likely encode unique functions. The predicted transmembrane domains in the conserved N- and C-terminal regions of each paralogue could serve as membrane anchors, with the variable regions forming extracellular loops involved in specific ligand binding, or other unique functions. The extensive 5′ and 3′ identities among the paralogues also could facilitate both intrastrain or interstrain recombination; interstrain recombination would be a powerful mechanism for increasing the functional repertoire of the recipient strains. That *babA* and *babB* are in opposite locations in relation to flanking genes in strains J99 and 26695 suggests that a reciprocal exchange could have occurred (8). The high (74 to 98%) level of identity between the downstream intergenic regions of HP317, *babA*, and *babB* and the downstream gene identity for HP317 and *babB* also could promote recombination.

Gene duplications among the *babA* paralogues have been observed for all three strains studied; strain CCUG17875 has two copies of *babA* (25), whereas strains 26695 and J99 have identical paralogues HP227 and HP1342 and JHP212 and JHP1261, respectively. Gene duplications potentially yield additional functional copies of the gene for enhanced expression of the product (14). The duplicate genes also may serve as the foci of gene conversion events that result in horizontal genetic movement between the copies (6, 7, 23). In strain CCUG17875, *babA2* (but not *babA1*) is necessary for Lewis B binding (25). Thus, *babA1* does not increase binding efficiency but may be useful for increasing *babA* diversity through gene conversion.

Segments of DNA, such as pathogenicity islands, that differ in GC content from the surrounding chromosome are believed to result from relatively recent cross-species acquisitions (29). That the *babA* paralogues have a consistently higher GC content (mean, 43%) than does the *H. pylori* genome (39%) suggests that they may have been a relatively recent genomic acquisition, followed by gene duplication events leading to the presence of multiple paralogues. The presence of HP317 in strain 26695 (absent in J99) could be explained by an even more recent gene duplication event or conversely by its deletion in J99. That the amount of intergenic similarity for each of the *babA* paralogues (90.4%) in 26695 and J99 is significantly less than that for the entire genomes (94.0%) suggests that, on average, they are diversifying faster than the rest of the *H. pylori* genome.

Allelic variation within individual *babA* paralogues has been previously reported for HP9 (*hopZ*; with two distinct alleles) and for HP1342 and HP227 (27, 36). For *hopZ* (36), the region of greatest diversity is located in nearly the analogous position to the *babA* and *babB* diversity regions. Despite the substantial polymorphism that exists throughout the *babA* fragments, the 42 strains that we studied cluster phylogenetically almost exactly according to the diversity present in the 84-nucleotide

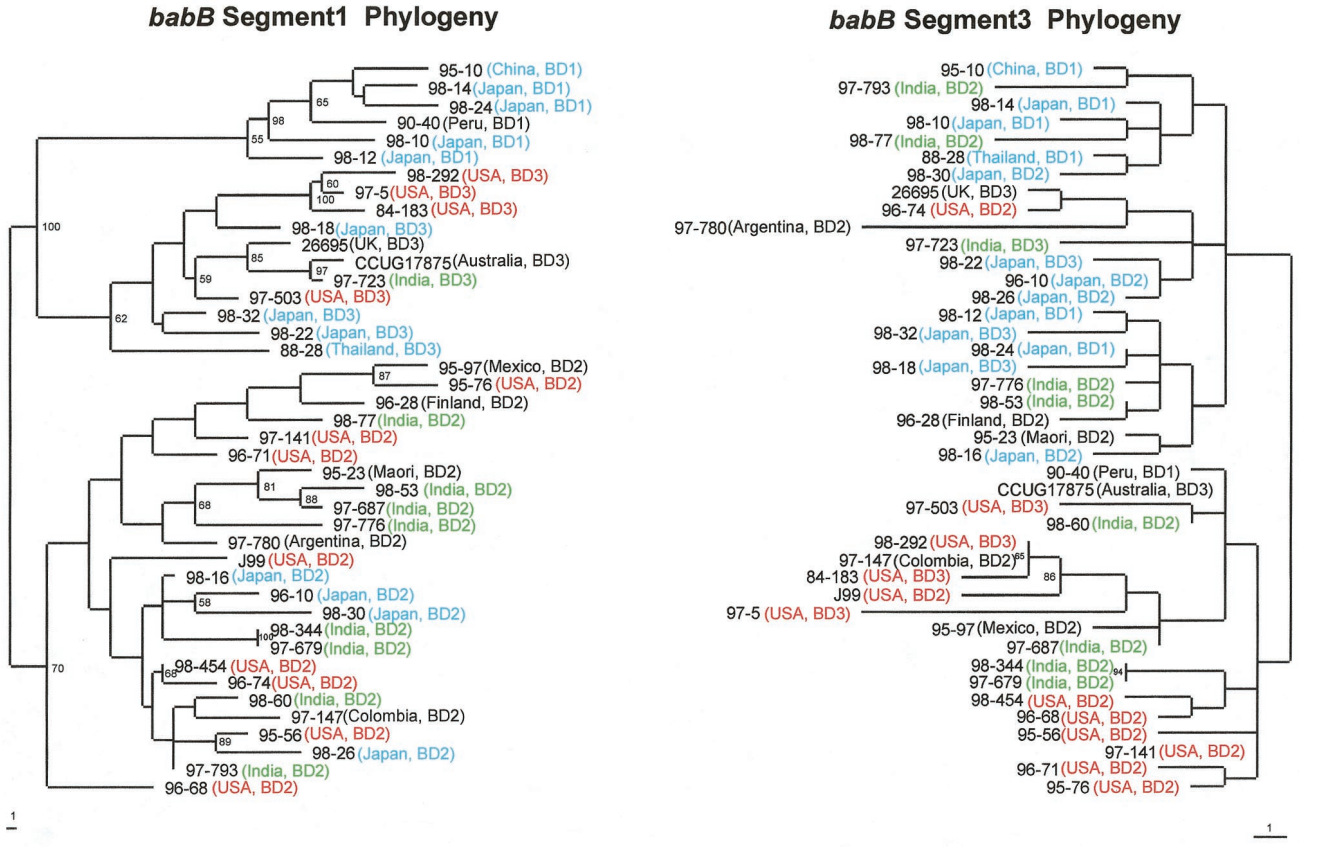

FIG. 6. Phylograms of *babB* segment 1 (nucleotides 646 to 754 on the left) and *babB* segment 3 (nucleotides 755 to 863 on the right). Segment 3 is the region downstream of segment 1, which is the region from which the alleles of *babB* were described. Fragments from 42 strains underwent sequence analysis as previously described and were aligned using GCG Pileup, and specific regions were identified using GCG Pretty. The sequences were then realigned using ClustalW and subjected to phylogenetic analysis using Paup 4.0b2 neighbor-joining analysis based on Kimura's two-parameter model distance matrices. Strain numbers, geographic locations, and allelic types are represented at the termination of each branch, and the branch length index is shown below each tree. Bootstrap values of >50 (based on 500 replicates) are shown at each node. Colors are used to represent strains from the United States (orange), Asia (not including India) (blue), India (green), and other countries (black).

allele group segment (data not shown); that this region dominates the phylogenetic structure suggests an important functional role. Strain 97-147 (AD5), which has features of each of the other *babA* allele groups, is most closely related to another South American strain (90-40, AD2). Its unique sequence indicates the possibility of recombination within the *babA* allele groups in the diversity region, paralleling (but less common than) recombination within the *babB* allele group region.

Most of the diversity in the *babB* fragments exists in a 108-nucleotide segment. Each of the different allele groups of *babB* contains greater diversity than that which exists within each *babA* allele group (Fig. 2, 3, and 5), which likely is due to interallelic recombination. In contrast to *babA*, phylogenetic analyses of the entire *babB* fragments show that strains do not segregate according to the allele groups but rather based on the geographic origin of each strain (data not shown). That the *babA* and *babB* allele group regions are largely independent of geographic origin but are flanked by regions that show evidence of geographic variation implies that these allele groups have moved horizontally throughout the *H. pylori* population. In support of this hypothesis, the similarity plot analysis indicates that the borders of each of the allele groups may represent recombination breakpoints (Fig. 3).

Overall, for the sequences studied, *babA* shows much more variation in length and lower average similarity than *babB*. In the third codon position, transversions are more likely to change amino acids in coding sequences, while transitions almost always leave coding sequences intact (Met and Trp are the only exceptions); thus, transitional substitutions tend to predominate for most species (30). In *H. pylori*, transitions account for most interstrain diversity, accounting for a mean of 80% (range, 66 to 94%) of the polymorphisms (45). However, in both *babA* and *babB* sequences, transitions account for only 50% of the polymorphisms; thus, transversions are far more common than the typical *H. pylori* gene (45). Because both *babA* and *babB* are outer membrane proteins, the diversity observed in this region in both genes may result from selection, possibly due to ongoing immune recognition.

Analysis of synonymous and nonsynonymous substitutions among the *babA* and *babB* diversity region fragments indicates that the *babB* fragments share a more recent common ancestor and that, as measured by the K_a/K_s ratio, both gene products are under similar functional constraints. That, for both genes, the K_a/K_s ratios for comparisons of sequences from specific allele groups are little different from one another (data not shown) and are considerably less than those observed for frag-

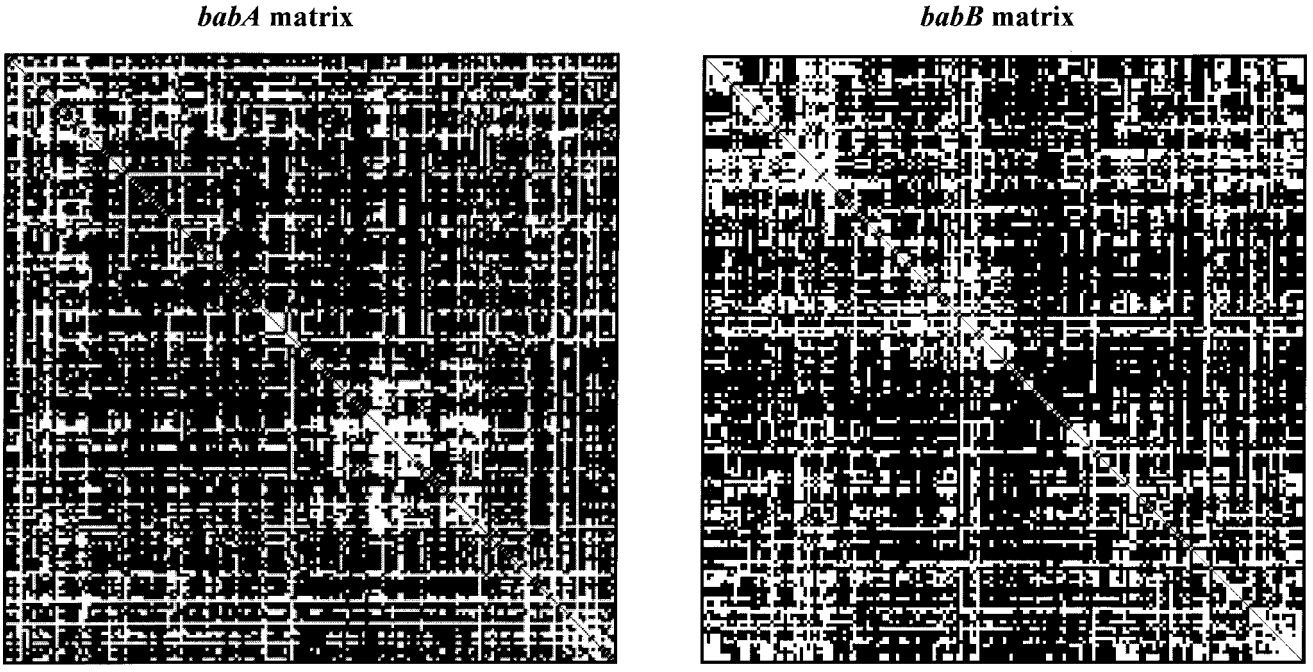

FIG. 7. Compatibility analysis of the *babA* (left) and the *babB* (right) diversity regions from 42 *H. pylori* strains. In the matrix, compatible sites are represented by white boxes and incompatible sites are represented as black boxes. Series of incompatible sites probably arose from repeated recombination, whereas compatible sites arose under clonal descent. Matrices were created using Reticulate (26).

ments within single alleles of either *cagA* and *vacA* indicates that there are similar functional restrictions between these allele groups (11, 42).

The existence of substantial recombination in both the *babA* and *babB* fragments is supported by the consistency index, compatibility matrices, and the homoplasy test. Homoplasies can arise through recombination or independent mutations, and the *H* ratio gives a measure of the observed synonymous homoplasies relative to the observed sequence variation (1, 33). The *H* ratios for *babA* and *babB* indicate that substantial recombination has likely occurred within both genes and at a level similar to that observed previously for several housekeeping genes (1). However, the presence of the allele groups and the geographic variation in both *babA* and *babB* (Fig. 4 and 6) indicate that recombination has not been sufficient to totally obscure evidence of clonal descent and suggest specific functional differences among the different allele groups. The high frequency of recombination for both *babA* and *babB* helps explain why the phylograms of the same strains are not congruent.

TABLE 5. Lewis B binding of J99 wild-type and J99 *babA2* and *babB* mutant strains

J99 genotype	Lewis B binding (relative OD units)[b,c]	
	Mean ± SD	Range
Wild type	653 ± 8	647–658
babA2[a]	8 ± 17	<10–20
babB[a]	596 ± 111	517–674

[a] Genes disrupted by insertion of *aphA* (Kan^r) cassette in specified ORF.
[b] Values represent the means of four separate experiments.
[c] Relative units are expressed as OD × 1,000 as described in Materials and Methods.

The presence of well-conserved allele groups in the *babA* and *babB* diversity regions implies an important functional role. *babA* has been shown previously to be responsible for Lewis B binding in *H. pylori* (25), which we now confirm. Because of the substantial similarity of *babB* and *babA, babB* might be involved in Lewis B binding as well; however, our data clearly indicate that neither the presence of *babB* allele groups nor the presence of *babA* allele groups is a determining factor in Lewis B binding.

In summary, both substantial conservation and variation exist among the *babA* paralogues. Two such paralogues, *babA* and *babB*, show both geographic and allelic group-associated variation in their predicted regions of maximum diversity. *babB* fragments appear to share a more recent common ancestor, but both the *babA* and *babB* gene products are under similar functional constraints. Although recombination accounts for much of the variation in *babA* and *babB*, as for *vacA*, it is not sufficient to obscure clonal structures present in both genes. Despite the involvement of *babA* in Lewis B binding, neither the *babA* allele groups nor the *babB* allele groups are determining factors in Lewis B binding. Whether the presence of different alleles of *babA* and *babB* has other functional implications could aid in our understanding of *H. pylori*-host ligand interactions.

ACKNOWLEDGMENTS

This work was supported in part by the Vanderbilt University Medical Scientist Training Program, the UNCF-Merck Science Initiative, and RO1 DK 53707 and the Cancer Center Core grant CA 68485 from the National Institutes of Health.

We thank Judith Romero and Margarita Carmolinga for the donation of DNA samples.

REFERENCES

1. **Achtman, M., T. Azuma, D. E. Berg, Y. Ito, G. Morelli, Z.-J. Pan, S. Suerbaum, S. A. Thompson, A. van der Ende, and L. van Doorn.** 1999. Recom-

bination and clonal groupings within *Helicobacter pylori* from different geographical regions. Mol. Microbiol. **32:**459–470.
2. **Akopyanz, N. S., N. O. Bukanov, T. U. Westblom, S. Kresovich, and D. E. Berg.** 1992. DNA diversity among clinical isolates of *Helicobacter pylori* detected by PCR-based RAPD fingerprinting. Nucleic Acids Res. **20:**5137–5142.
3. **Akopyanz, N. S., N. O. Bukanov, T. U. Westblom, and D. E. Berg.** 1992. PCR-based RFLP analysis of DNA sequence diversity in the gastric pathogen *Helicobacter pylori*. Nucleic Acids Res. **20:**6221–6225.
4. **Akopyants, N. S., S. W. Clifton, D. Kersulyte, J. E. Crabtree, B. E. Youree, C. A. Reece, N. O. Bukanov, E. S. Drazek, B. A. Roe, and D. E. Berg.** 1998. Analyses of the *cag* pathogenicity island of *Helicobacter pylori*. Mol. Microbiol. **28:**37–54.
5. **Alm, R. A., L.-S. L. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, B. Noonan, B. C. Guild, B. L. deJonge, G. Carmel, P. J. Tummino, A. Caruso, M. Uria-Nickelsen, D. M. Mills, C. Ives, R. Gibson, D. Merberg, S. D. Mills, Q. Jiang, D. E. Taylor, G. F. Vovis, and T. J. Trust.** 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature **397:**176–180.
6. **Alm, R. A., P. Guerry, and T. J. Trust.** 1993. Distribution and polymorphism of the flagellin genes from isolates of *Campylobacter coli* and *Campylobacter jejuni*. J. Bacteriol. **175:** 3051–3057.
7. **Alm, R. A., P. Guerry, and T. J. Trust.** 1993. Significance of duplicated genes in *Campylobacter*. J. Mol. Biol. **230:**359–363.
8. **Alm, R. A., J. Bina, B. M. Andrews, P. Doig, R. E. W. Hancock, and T. J. Trust.** 2000. Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. Infect. Immun. **68:**4155–4168.
9. **Altschul, S. F., Y. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
10. **Atherton, J. C., P. Cao, R. M. J. Peek, M. K. Tummuru, M. J. Blaser, and T. L. Cover.** 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.
11. **Atherton, J. C., P. M. Sharp, T. L. Cover, G. Gonzalez-Valencia, R. M. Peek, S. A. Thompson, C. J. Hawkey, and M. J. Blaser.** 1999. 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.
12. **Blaser, M. J., and J. Parsonet.** 1994. Parasitism by the "slow" bacterium *Helicobacter pylori* leads to altered gastric homeostasis and neoplasia. J. Clin. Investig. **94:**4–8.
13. **Boren, T., P. Falk, K. A. Roth, G. Larson, and S. Normark.** 1993. Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. Science **262:**1892–1895.
14. **Brown, C. J., K. M. Todd, and R. F. Rosenzweig.** 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol. Biol. Evol. **15:**931–942.
15. **Censini, S., C. Lange, Z. Xiang, J. E. Crabtree, P. Ghiara, M. Borodovsky, R. Rappuoli, and A. Cocacci.** 1996. *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. Proc. Natl. Acad. Sci. USA **93:**14648–14653.
16. **Covacci, A., S. Censini, M. Bugnoli, R. Petracca, D. Burroni, G. Macchia, A. Massone, E. Papini, Z. Xiang, and N. Figura.** 1993. Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. Proc. Natl. Acad. Sci. USA **90:**5791–5795.
17. **Cover, T. L., M. K. Tummuru, P. Cao, S. A. Thompson, and M. J. Blaser.** 1994. Divergence of genetic sequences for the vacuolating cytotoxin among Helicobacter pylori strains. J. Biol. Chem. **269:**10566–10573.
18. **Doig, P., M. M. Exner, R. E. W. Hancock, and T. J. Trust.** 1995. Isolation and characterization of a conserved porin protein from *Helicobacter pylori*. J. Bacteriol. **177:**5447–5452.
19. **Exner, M. M., P. Doig, T. J. Trust, and R. E. W. Hancock.** 1995. Isolation and characterization of a family of porin proteins from *Helicobacter pylori*. Infect. Immun. **63:**1567–1572.
20. **Fujimoto, S., B. Marshall, and M. J. Blaser.** 1994. PCR-based restriction fragment polymorphism typing of *Helicobacter pylori*. J. Clin. Microbiol. **32:**331–334.
21. **Garnier, J., D. J. Osguthorpe, and B. Robson.** 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120:**97–120.
22. **Go, M. F., V. Kapur, D. Y. Graham, and J. M. Musser.** 1996. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. J. Bacteriol. **178:**3934–3938.
23. **Guerry, P. R., R. A. Alm, M. E. Power, S. M. Logan, and T. J. Trust.** 1991. Role of two flagellin genes in *Campylobacter* motility. J. Bacteriol. **173:**4757–4764.
24. **Hofmann, K., and W. Stoffel.** 1993. TMbase—a database of membrane spanning protein segments. Biol. Chem. Hoppe-Seyler **347:**166.
25. **Ilver, D., A. Arnqvist, J. Ögren, I.-M. Frick, D. Kersulyte, E. T. Incecik, D. E. Berg, A. Covacci, L. Engstrand, and T. Borèn.** 1998. *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. Science **279:**373–377.
26. **Jacobson, I. B, and S. Easteal.** 1996. A program for calculating matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:**291–295.
27. **Kersulyte, D., H. Chalkauskas, and D. E. Berg.** 1999. Emergence of recombinant strains of *Helicobacter pylori* during infection. Mol. Microbiol. **31:**31–43.
28. **Kuipers, E. J., G. I. Perez-Perez, S. G. Meuwissen, and M. J. Blaser.** 1995. *Helicobacter pylori* and atrophic gastritis: importance of the *cagA* status. J. Natl. Cancer Inst. **87:**1777–1780.
29. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:**383–397.
30. **Li, W.-H.** 1997. Molecular evolution, p. 309–334. Sinauer Associates, Inc., Publishers, Sunderland, Mass.
31. **Logan, R. P. H., and D. E. Berg.** 1996. Genetic diversity of *Helicobacter pylori*. Lancet **348:**1462–1463.
32. **Marshall, D. G., D. C. Coleman, D. J. Sullivan, H. Xia, C. A. O'Morain, and C. J. Smyth.** 1996. Genomic DNA fingerprinting of clinical isolates of *Helicobacter pylori* using short oligonucleotide probes containing repetitive sequences. J. Appl. Bacteriol. **81:**509–517.
33. **Maynard Smith, J., and N. H. Smith.** 1998. Detecting recombination from gene trees. Mol. Biol. Evol. **15:**590–599.
34. **Odenbreit, S., M. Till, D. Hofreuter, G. Faller, and R. Haas.** 1999. Genetic and functional characterization of the *AlpAB* gene locus essential for the adhesion of *Helicobacter pylori* to human gastric tissue. Mol. Microbiol. **31:**1537–1548.
35. **Page, R. D. M.** 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:**357–358.
36. **Peck, B., M. Ortkamp, K. D. Diehl, E. Hundt, and B. Knapp.** 1999. Conservation, localization, and expression of *hopZ*, a protein involved in adhesion of *Helicobacter pylori*. Nucleic Acids Res. **27:**3325–3333.
37. **Rozas, J., and R. Rozas.** 1999. DnaSP version 3.0: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:**174–175.
38. **Suerbaum, S., J. M. Smith, K. Bapumia, G. Morelli, N. H. Smith, E. Kunstmann, I. Dyrek, and M. Achtman.** 1998. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA **95:**12619–12624.
39. **Taylor, N. S., J. G. Fox, N. S. Akopyants, D. E. Berg, N. Thompson, B. Shames, L. Yan, E. Fontham, F. Janney, F. M. Hunter, and P. Correa.** 1995. Long-term colonization with single and multiple strains of *Helicobacter pylori* assessed by DNA fingerprinting. J. Clin. Microbiol. **33:**918–923.
40. **Tomb, J.-F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischman, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Xhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzgerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, M. D. Cotton, J. M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter.** 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature **388:**539–547.
41. **Tummuru, M. K., T. L. Cover, and M. J. Blaser.** 1993. Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori*: evidence of linkage to cytotoxin production. Infect. Immun. **61:**1799–1809.
42. **van Doorn, L.-J., C. Figueiredo, R. Sanna, M. J. Blaser, and W. G. V. Quint.** 1999. Distinct variants of *Helicobacter pylori cagA* are associated with *vacA* subtypes. J. Clin. Microbiol. **37:**2306–2311.
43. **van Doorn, L.-J., C. Figueiredo, R. Sanna, S. Pena, P. Midolo, E. K. W. Ng, J. C. Atherton, M. J. Blaser, and W. G. V. Quint.** 1998. Expanding allelic diversity of *Helicobacter pylori vacA*. J. Clin. Microbiol. **36:**2597–2603.
44. **van Doorn, L. J., C. Figueiredo, R. Sanna, A. Plaisier, P. Schneeberger, W. D. Boer, and W. G. V. Quint.** 1998. Clinical relevance of the *cagA, vacA*, and *iceA* status of *Helicobacter pylori*. Gastroenterology **115:**58–66.
45. **Wang, G. E., M. Z. Humayun, and D. E. Taylor.** 1999. Mutation as an origin of genetic variability in *Helicobacter pylori*. Trends Microbiol. **7:**488–493.
46. **Wilkins, M. R., E. Gasteiger, A. Bairoch, J.-C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser.** 1998. Protein identification and analysis in the ExPaSy Server, p. 531–552. *In* A. J. Link (ed.), 2-D proteome analysis protocols. Humana Press, Inc., Totowa, N.J.
47. **Wilson, K.** 1995. Preparation of genomic DNA from bacteria, p. 2.4.1-2.4.5. *In* F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (ed.), Current protocols in molecular biology, vol. 1. John Wiley & Sons, Inc., New York, N.Y.
48. **Yamaoka, Y., T. Kodama, O. Gutierrez, J. G. Kim, K. Kashima, and D. Y. Graham.** 1999. Relationship between *Helicobacter pylori iceA, cagA*, and *vacA* status and clinical outcome: studies in four different countries. J. Clin. Microbiol. **37:**2274–2279.

Editor: A. D. O'Brien