Genomic Comparison of *cag* Pathogenicity Island (PAI)-Positive and -Negative *Helicobacter pylori* Strains: Identification of Novel Markers for *cag* PAI-Positive Strains

Courtney E. Terry,[1]† Lisa M. McGinnis,[1] Katharine C. Madigan,[1]‡ Ping Cao,[2] Timothy L. Cover,[2,4]
George W. Liechti,[1] Richard M. Peek, Jr.,[3,4] and Mark H. Forsyth[1]*

*The College of William and Mary, Department of Biology, Williamsburg, Virginia 23187-8795[1]; Vanderbilt University
School of Medicine, Division of Infectious Diseases, Departments of Medicine and Microbiology
and Immunology, Nashville, Tennessee 37232-2279[2]; Vanderbilt University School of
Medicine, Department of Medicine, Division of Gastroenterology, and Department
of Cancer Biology, Nashville, Tennessee 37232-2279[3]; and
Department of Veterans Affairs Medical Center,
Nashville, Tennessee 37212[4]*

In an analysis of *Helicobacter pylori* genomic DNA by macroarray methodology, genomic DNA from a panel of *cag* pathogenicity island (PAI)-negative *H. pylori* clinical isolates failed to hybridize with 27 genes located outside the *cag* PAI in a *cag* PAI-positive reference strain. PCR analyses confirmed that HP0217 (encoding a lipopolysaccharide biosynthetic protein) and HP1079 (encoding a protein of unknown function) were present significantly more frequently in *cagA*-positive strains than in *cagA*-negative strains. A low G+C content of these two genes suggests they were acquired by horizontal transfer events.

Helicobacter pylori is recognized as the major etiologic agent of peptic ulcer disease and gastric neoplasia (5, 6). This gram-negative microaerophile exhibits tremendous genetic diversity (9, 10) due to a combination of factors, including the organism's high mutation rate (2, 15), its natural competence for uptake of foreign DNA (4, 8), its ability to undergo frequent homologous recombination (12, 13), its evolution in geographically restricted environments (16), and an ancient evolutionary history (7). One potential consequence of this genetic diversity may be variation in disease outcome among infected individuals.

A major genetic determinant of *H. pylori* virulence is the *cag* pathogenicity island (*cag* PAI) (3, 5, 10), a 40-kb region of chromosomal DNA that is present in some *H. pylori* strains but absent from others. The *cag* PAI encodes a type IV secretion system and an immunodominant antigen, CagA, which is translocated into gastric epithelial cells. In comparison to infection with *cag* PAI-negative *H. pylori* strains, infection with *cag* PAI-positive strains is associated with an increased severity of gastric mucosal inflammation, an increased risk for development of peptic ulceration, and an increased risk of gastric cancer (3).

The complete genomes of two *cag* PAI-positive strains of *H. pylori* (26695 and J99) have been sequenced (1, 14). Despite similarity at two major disease-associated loci (both are *cag* PAI positive and contain type s1 *vacA* alleles), strain 26695

(14) contains 110 open reading frames (ORFs) not found in strain J99 (1) and strain J99 contains 52 genes that are not found in strain 26695. In a comprehensive examination of *H. pylori* genetic diversity, Salama et al. (11) identified 362 *H. pylori* genes that were each absent in at least one of 15 strains examined and suggested that the core genome of *H. pylori* consists of approximately 1,300 genes. These data suggest that insertion and deletion of sequences occur commonly in *H. pylori*. We hypothesized that there may be differential retention of specific genetic elements that are advantageous for *cag* PAI-positive organisms in an inflammatory gastric mucosal environment or deletion of genetic elements that are disadvantageous for *cag* PAI-negative organisms. Thus, in the current study, we sought to identify genes that are present more frequently in *cag* PAI-positive strains than in *cag* PAI-negative strains.

To identify such genes, we selected five *H. pylori* isolates that were genetically characterized as *cagA* negative and *vacA* s2/m2 for use in DNA macroarray analyses. (Detailed strain information is available upon request.) These are genotypic markers for *H. pylori* strains that are associated with a low risk for the development of clinical disease. At the time of endoscopy, each of the five source patients was diagnosed with gastritis only, and none of these patients had a prior history of peptic ulcer disease.

Our analyses utilized DNA macroarrays (Sigma-Genosys) containing 1,681 known *H. pylori* ORFs found in the genomes of two *cag* PAI-positive sequenced strains of *H. pylori*. Arrays were individually hybridized with ^{33}P-labeled genomic DNA from the five *cagA*-negative isolates and DNA from a *cag* PAI-positive sequenced strain (26695) as a control. *H. pylori* genomic DNA was labeled with ^{33}P using random-primed DNA labeling (Promega) and [α-^{33}P]dCTP (Perkin-Elmer).

* Corresponding author. Mailing address: Department of Biology, The College of William & Mary, 214 Millington Hall, Landrum Drive, Williamsburg, VA 23187-8795. Phone: (757) 221-2489. Fax: (757) 221-6483. E-mail: mhfors@wm.edu.
† Present address: University of Virginia School of Medicine, Charlottesville, VA 22908.
‡ Present address: Thomas Jefferson University, Jefferson Medical College, Philadelphia, PA 19107.

TABLE 1. *H. pylori* genes absent from a set of five *cagA*-negative clinical isolates as determined by macroarray analyses[a]

H. pylori strain 26695 designation	*H. pylori* strain J99 designation	Putative function
HP0053	N/A	Type II restriction enzyme
HP0217	*jhp0203*	LPS biosynthesis, β-1,4-*N*-acetylgalactosamyltransferase (phase variable)
HP0260	*jhp0244*	Type III restriction enzyme
HP0336	N/A	Cysteine rich protein B, peptidoglycan synthesis
HP0433	N/A	*H. pylori*-specific hypothetical protein
HP0434	N/A	*H. pylori*-specific hypothetical protein
HP0435	N/A	*H. pylori*-specific hypothetical protein
HP0437	N/A	IS*605* transposase A
HP0438	N/A	IS*605* transposase B
HP0448	N/A	*H. pylori*-specific hypothetical protein
HP0449	N/A	*H. pylori*-specific hypothetical protein
HP0452	N/A	*H. pylori*-specific hypothetical protein
HP0461	N/A	*H. pylori*-specific hypothetical protein
HP0882	N/A	*H. pylori*-specific hypothetical protein
HP0892	*jhp0831*	Conserved metal-dependent enzyme
HP0893	*jhp0832*	*H. pylori*-specific hypothetical protein
HP0998	N/A	IS*605* transposase A
HP1079	*jhp0346*	*H. pylori*-specific hypothetical protein
HP1096	N/A	IS*605* transposase A
HP1097	N/A	*H. pylori*-specific hypothetical protein
HP1142	*jhp1070*	*H. pylori*-specific hypothetical protein
HP1209	N/A	Type II restriction enzyme
HP1390	N/A	*H. pylori*-specific hypothetical protein
HP1437	*jhp1330*	*H. pylori*-specific hypothetical protein
HP1438	*jhp1331*	*H. pylori*-specific hypothetical protein
HP1535	N/A	IS*605* transposase A
HP1578	N/A	LPS biosynthesis

[a] Numerous genes of the *cag* PAI were absent in all five of the assayed *cagA*-negative *H. pylori* clinical isolates. Those genes are not listed in this table.

Arrays were imaged using a Storm 840 PhosphorImager (Molecular Dynamics) and signals quantified using ArrayVision (Imaging Research, St. Catharines, Ontario, Canada). Background hybridization was quantified based on analysis of 45 macroarray features on which no DNA was arrayed. The mean background value was subtracted from values for all other array features. Data from individual arrays were normalized by expressing the value from each array feature as a percentage of the total signal for the entire macroarray.

To identify genes absent from the five *cagA*-negative query strains, we compared the array results obtained with DNA from *cagA*-negative strains with the array results obtained with DNA from the *cag* PAI-positive reference strain. For each array feature, a ratio was calculated by dividing the normalized signal intensity for a *cagA*-negative strain by the corresponding normalized signal intensity value obtained with the *H. pylori* *cag* PAI-positive reference strain 26695. Genes whose features yielded ratios of ≤0.2 were considered absent in the tested *cagA*-negative clinical isolate.

DNA from each of the five *cagA*-negative clinical isolates failed to hybridize with multiple array features (mean, 109 features; range, 61 to 180), including genes comprising the *cag* PAI (data not shown). All five *cagA*-negative strains failed to hybridize with 27 genes located outside the *cag* PAI in the chromosome of the *H. pylori* 26695 reference strain (Table 1). Nine of these 27 genes (HP0433 to HP0461) have been mapped to a region of the *H. pylori* chromosome known as the plasticity zone, a ~44-kb region that is enriched in strain-specific *H. pylori* genes (1). Three of the 27 genes encode products that are predicted to be involved in DNA restriction/modification, five are involved in DNA transposition, 15 en-

code *H. pylori*-specific proteins of unknown function, and four encode proteins with various other predicted functions (Table 1).

We next used PCR-based assays to test for the presence or absence of three of these genes (HP0217, HP1079, and HP1578) in a set of 18 *cagA*-positive and 14 *cagA*-negative *H. pylori* clinical isolates (independent of the five *cag*-PAI negative strains utilized in the initial DNA macroarray studies described above) (Table 2). All 18 *cagA*-positive strains selected for study contained a type s1 *vacA* allele, and all 14 *cagA*-negative *H. pylori* isolates contained a type s2/m2 *vacA* allele. Among the three genes selected for analysis by PCR, two (HP0217 and HP1578) are predicted to be involved in lipo-

TABLE 2. PCR analyses of gene frequencies among a set of 32 clinical isolates of *H. pylori*

Gene designation[a]	*cagA* positive[b]	*cagA* negative[b] (P)
HP0216	100	100
HP0217	89	7 (<0.001)
HP0218	100	100
HP1077	94	100
HP1079	72	21 (<0.05)
HP1080	89	100
HP1577	94	100
HP1578	22	0
HP1579	83	100

[a] Gene designations based upon *H. pylori* strain 26695 genome annotation (14).
[b] Data given as the percent of isolates yielding amplicons of the expected size in PCR analyses, using gene-specific primers. Positive, *n* = 18; negative, *n* = 14.
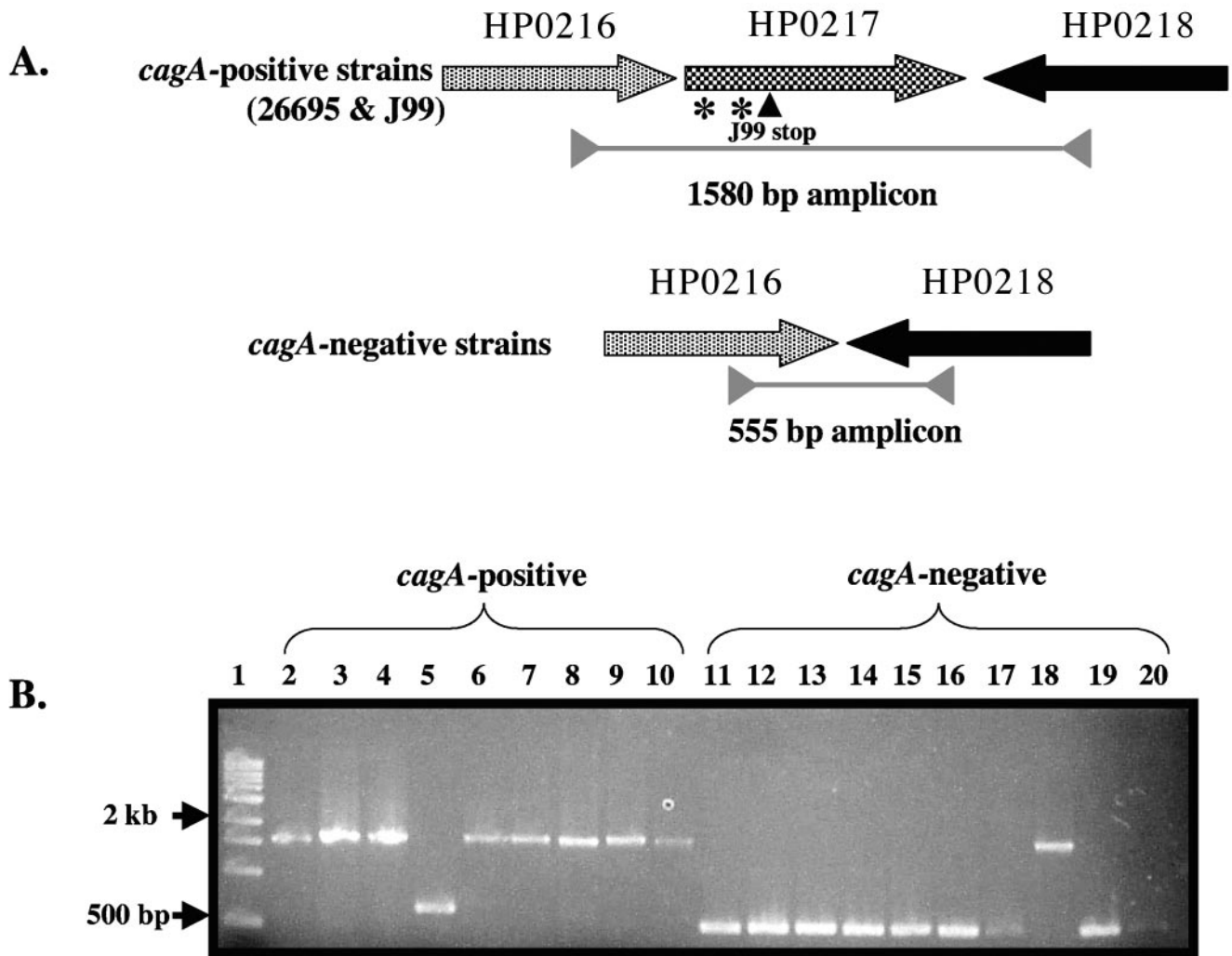
FIG. 1. HP0217 is more frequent among *cagA*-positive *H. pylori* isolates than among *cagA-negative* isolates. A. The genomic organization of the variable locus HP0217 in the sequenced *cag* PAI-positive strains, 26695 and J99, is shown in comparison to the most common deduced arrangement of this locus in *cagA*-negative *H. pylori* isolates. Gray triangles represent the location of the empty-site primers used in PCR analysis (see below). ∗, denotes location of homopolymeric repeats (poly G) where phase variation may occur. HP0217 is predicted to encode a full-length *N*-acetylgalactosamyltransferase in *H. pylori* strain 26695 and a truncated protein in strain J99. B. Empty-site primers derived from HP0216 and HP0218 (5′ GCTGAATGCGAGCAATGAAGTGGCG 3′ and 5′ GCGTCTATATCGGCCCCATGCC) were used to amplify intervening sequences from *cagA*-positive and *cagA*-negative *H. pylori* strains. Amplicons consistent with the presence of HP0217 (∼1,500 bp) were amplified more frequently from *cagA*-positive isolates than *cagA*-negative isolates. A 555-bp amplicon was more frequently amplified from the latter strains. A representative sample of the *H. pylori* strains used in this study is shown here. *H. pylori cagA*-positive isolates, lanes 2–10; *cagA*-negative isolates, lanes 11–20. Lane 1, 1-kb ladder.

polysaccharide (LPS) biosynthesis, and one (HP1079) encodes an *H. pylori*-specific product of unknown function.

To investigate the presence or absence of HP0217 in this group of 32 *H. pylori* strains, we designed sets of primers for PCR amplification of HP0217 and two flanking genes (HP0216 and HP0218). (All primer sequences are available upon request.) PCR analyses indicated that both flanking genes were present in all 32 *H. pylori* strains examined, regardless of *cagA* status (Table 2). In contrast, HP0217 sequences were amplified from only 53% (17 of 32) of strains. HP0217 sequences were successfully amplified from 89% (16 of 18) of *cagA*-positive strains but from only one (7%) of the 14 *cagA*-negative isolates examined ($P < 0.001$) (Table 2).

We also performed a second PCR analysis, using primers designed to anneal within the conserved HP0216 and HP0218 genes (Fig. 1A). This empty-site PCR analysis was predicted to yield a ∼1.5-kb amplicon if HP0217 was present and a 0.5-kb amplicon if HP0217 was absent. Most (14/18) of the *cagA*-positive strains yielded a 1.5-kb amplicon. Thirteen of the 14 *cagA*-negative strains yielded a 0.5-kb amplicon, and one (strain 92-24) yielded a 1.5-kb amplicon (Fig. 1B). Sequence analysis of the 555-bp amplicon from a representative *cagA*-negative strain (*H. pylori* 92-28) confirmed the absence of any portion of HP0217 and revealed the presence of a 180-bp segment (GenBank accession no. AY529682) that had no significant homology to sequences in either of the sequenced strains of *H. pylori*, 26695 or J99. Thus, two different PCR assays indicated that HP0217 is found in *cagA*-positive *H. pylori*

strains significantly more frequently than in *cagA*-negative *H. pylori* strains ($P < 0.001$).

The gene product of HP0217 is predicted to play a role in LPS biosynthesis and is predicted to undergo phase variation based on the presence of two poly-G tracts within the 5′ portion of the gene (Fig. 1A). These poly-G tracts may be substrates for slipped-strand mispairing events, leading to frameshift mutations. Based on the demonstrated differential distribution of HP0217 in *cag* PAI-positive and *cag* PAI-negative strains of *H. pylori*, we hypothesize that there may be differences in LPS oligosaccharides of *cagA*-positive and *cagA*-negative strains. In fact, it has been reported that *cagA*-positive strains of *H. pylori* express greater amounts of LPS-associated Lewis antigens on their surface than do *cagA*-negative strains (17).

We analyzed the presence or absence of HP1079 in the collection of 32 *H. pylori* clinical isolates using methodology similar to that described for HP0217. HP1080 and HP1077 sequences were amplified from nearly all strains tested (Table 2). HP1079 sequences were amplified from 13 of 18 (72%) *cagA*-positive strains but from only 3 of 14 (21%) *cagA*-negative strains examined ($P < 0.05$) (Table 2). In an HP1080 to HP1077 empty-site PCR analysis (primer sequences available upon request), 14 (78%) of the 18 *cag* PAI-positive strains yielded a ~2.4-kb amplicon, indicating the presence of orthologs of HP1079 and HP1078 in these 14 *cagA*-positive strains. One (5.5%) of the 18 *cagA*-positive strains yielded a small ~500-bp amplicon in this empty-site PCR analysis, suggesting the absence of orthologs of HP1079 and HP1078 in this strain, and no product was amplified from three of the *cagA*-positive strains. Thirteen of 14 (93%) of the *cagA*-negative strains yielded a small ~500-bp amplicon, suggesting the absence of HP1079/HP1078 orthologs in these 13 strains. The sequence of this amplicon from a representative *cagA*-negative *H. pylori* strain (92-28) confirmed the absence of any coding sequence in the region between HP1077 and HP1080 (GenBank accession no. AY529680). One *cagA*-negative strain that yielded a 2.5-kb amplicon in this assay was demonstrated to possess HP1079, based on PCR analysis using HP1079 gene-specific primers. Two *cag* PAI-negative strains (87-75 and J195) yielded 0.5-kb amplicons in the empty-site PCR analysis but were demonstrated to contain HP1079 based on the PCR using HP1079 gene-specific primers. Potentially these two strains possess HP1079 at another site in the genome. In summary, based on the results of two different PCR assays, HP1079 sequences were detected significantly more commonly in *cagA*-positive *H. pylori* strains than in *cagA*-negative strains.

HP1578 was found to be a relatively rare gene in *H. pylori*. It was amplified from only 4 of 32 clinical isolates, each of which was also *cag* PAI positive. The difference in prevalence of HP1578 in *cagA*-positive and *cagA*-negative isolates was not statistically significant (Table 2).

One potential explanation for the presence of genes in some *H. pylori* strains and not in others is that these genes may have been acquired via horizontal transfer events. While the overall reported G+C content for both sequenced *H. pylori* strains is 39% (1, 14), the G+C content of HP1079 is only 29.6%. The G+C contents of the conserved genes flanking the HP1079 locus are 37.7% and 39.8%, respectively. Similarly, the G+C content of HP0217 is relatively low (33.9%), whereas the G+C

contents of the two ORFs upstream of HP0217 and the two ORFs downstream of HP0217 are 42.9% and 40.4%, respectively. The low G+C content of HP1079 and HP0217 suggests that these genes were acquired via horizontal transfer events.

The results of this study may be compared and contrasted with a *H. pylori* genomic analysis published previously by Salama et al. (11). Salama et al. used microarrays to analyze the genomic content of 15 *H. pylori* strains (11 *cagA* positive and 4 *cagA* negative) and identified 10 genes located outside of the *cag* PAI that were present significantly more frequently in *cagA*-positive strains than in *cagA*-negative strains. It was suggested that these genes may encode undescribed virulence factors. Notably, the two genes characterized in detail in the current study (HP0217 and HP1079) were not noted to covary with the *cag* PAI in the study by Salama et al. In both the current study (Table 1) and that of Salama et al., HP0260 (encoding a restriction enzyme) was found more frequently in *cagA*-positive strains than in *cagA*-negative strains. Genes such as HP0217, HP0053, HP0336, HP1142, and HP1578 were found infrequently in *cagA*-negative isolates of *H. pylori* in both studies but were not previously denoted as covarying with the *cag* PAI because of the absence of these genes in some *cag* PAI-positive isolates. Two of the *cag* PAI-covarying genes identified by Salama et al., HP1243 (*babA*) and HP1417, were absent in four of the five *cag* PAI-negative clinical isolates analyzed by macroarray analysis in the current study (data not shown). Although based on similar DNA array methodology, our study differed from that of Salama et al. (11) by utilizing exclusively *cag* PAI-negative *H. pylori* isolates with a type s2/m2 *vacA* genotype that were isolated from patients with superficial gastritis only. Therefore, some of the differences among the lists of *cag* PAI-covarying genes generated in these two studies are potentially attributable to differences in the clinical status of source patients or differences in the characteristics of strains selected for study. The results of the current study extend our understanding of genes existing in linkage disequilibrium with the *cag* PAI. Furthermore, several genes identified here may represent useful markers for the identification of virulent strains or may represent novel virulence factors.

H. pylori is a panmictic species (12, 13), and it has been suggested that *cag* PAI-positive *H. pylori* isolates are no more closely related to one another than they are to strains that lack this PAI (11). In light of the recombinatorial structure of the *H. pylori* genome, it is striking that there appears to be a selective pressure for *H. pylori* strains possessing the *cag* PAI to retain multiple strain-specific genes (e.g., HP0217 and HP1079) located elsewhere in the chromosome. Alternatively, there might be selective pressure on *H. pylori* strains lacking the *cag* PAI to delete these genes. In fact, both pressures may act simultaneously. We speculate that the severe mucosal inflammation associated with *cag* PAI-positive strains may represent one important environmental variable that serves as a powerful selective force.

REFERENCES

1. **Alm, R. A., L. S. 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.

2. **Bjorkholm, B., M. Sjolund., P. G. Falk, O. G. Berg, L. Engstrand, and D. I. Andersson.** 2001. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA **98:**14607–14612.

3. **Blaser, M. J., G. I. Perez-Perez, H. Kleanthous, T. L. Cover, R. M. Peek, Jr., P. H. Chyou, G. N. Stemmermann, and A. Nomura.** 1995. Infection with *Helicobacter pylori* strains possessing *cagA* is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res. **55:**2111–2115.

4. **Chang, K. C., Y. C. Yeh, T. L. Lin, and J. T. Wang.** 2001. Identification of genes associated with natural competence in *Helicobacter pylori* by transposon shuttle random mutagenesis. Biochem. Biophys. Res. Commun. **288:**961–968.

5. **Cover, T. L., D. E. Berg, M. J. Blaser, and H. L. T. Mobley.** 2001. *Helicobacter pylori* pathogenesis. Academic Press, San Diego, Calif.

6. **Dunn, B. E., H. Cohen, and M. J. Blaser.** 1997. *Helicobacter pylori*. Clin. Microbiol. Rev. **10:**720–741.

7. **Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Megraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum.** 2003. Traces of human migrations in *Helicobacter pylori* populations. Science **299:**1582–1585.

8. **Hofreuter, D., S. Odenbreit, and R. Haas.** 2001. Natural transformation competence in *Helicobacter pylori* is mediated by the basic components of a type IV secretion system. Mol. Microbiol. **41:**379–391.

9. **Labigne, A., and H. de Reuse.** 1996. Determinants of *Helicobacter pylori* pathogenicity. Infect. Agents Dis. **5:**191–202.

10. **Mobley, H. L.** 1996. Defining *Helicobacter pylori* as a pathogen: strain heterogeneity and virulence. Am. J. Med. **100:**2S–11S.

11. **Salama, N., K. Guillemin, T. K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow.** 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. USA **97:**14668–14673.

12. **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.

13. **Suerbaum, S., and M. Achtman.** 1999. Evolution of *Helicobacter pylori*: the role of recombination. Trends Microbiol. **7:**182–184.

14. **Tomb, J. F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, J. C. Venter, et al.** 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature **388:**539–547.

15. **Wang, G., M. Z. Humayun, and D. E. Taylor.** 1999. Mutation as an origin of genetic variability in *Helicobacter pylori*. Trends Microbiol. **7:**488–493.

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

17. **Wirth, H. P., M. Yang, M. Karita, and M. J. Blaser.** 1996. Expression of the human cell surface glycoconjugates Lewis X and Lewis Y by *Helicobacter pylori* isolates is related to *cagA* status. Infect. Immun. **64:**4598–4605.