Comparative genomics of the restriction-modification systems in *Helicobacter pylori*

Lee-Fong Lin, Janos Posfai, Richard J. Roberts, and Huimin Kong*

New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915

Communicated by Charles R. Cantor, Sequenom Industrial Genomics, Inc., San Diego, CA, December 22, 2000 (received for review October 31, 2000)

Helicobacter pylori **is a Gram-negative bacterial pathogen with a small genome of 1.64–1.67 Mb. More than 20 putative DNA restriction-modification (R-M) systems, comprising more than 4% of the total genome, have been identified in the two completely sequenced *H. pylori* strains, 26695 and J99, based on sequence similarities. In this study, we have investigated the biochemical activities of 14 Type II R-M systems in *H. pylori* 26695. Less than 30% of the Type II R-M systems in 26695 are fully functional, similar to the results obtained from strain J99. Although nearly 90% of the R-M genes are shared by the two *H. pylori* strains, different sets of these R-M genes are functionally active in each strain. Interestingly, all strain-specific R-M genes are active, whereas most shared genes are inactive. This agrees with the notion that strain-specific genes have been acquired more recently through horizontal transfer from other bacteria and selected for function. Thus, they are less likely to be impaired by random mutations. Our results also show that *H. pylori* has extremely diversified R-M systems in different strains, and that the diversity may be maintained by constantly acquiring new R-M systems and by inactivating and deleting the old ones.**

Helicobacter pylori is one of the most common bacterial pathogens that colonizes the gastric mucosa of humans. *H. pylori* is implicated in a wide range of gastroduodenal diseases (1, 2). *H. pylori* is commonly believed to be a very diverse species. It is believed that, in addition to genetic recombination, *de novo* mutation could have a role in generating the high level of genetic variation in *H. pylori* (3). MutH and MutL homologues cannot be found in *H. pylori* genomes, which suggests *H. pylori* may not have a functional mismatch repair system (4, 5). Recent analysis of the complete genomic sequences of two unrelated *H. pylori* isolates reveals that although intraspecies variation does exist, the overall genomic organization, gene order, and predicted proteins of the two strains are quite similar (5, 6). Approximately 6–7% of the genes are specific to each strain (5). The 26695 and J99 strains have a relatively small genome size of 1.67 and 1.64 megabase pairs (4, 5). However, more than twenty DNA restriction-modification (R-M) systems can be identified in each strain based on sequence similarities. The biological significance of this large complement of R-M systems is not clear. The majority of the *H. pylori* R-M systems are of Type II, which consist of two separate enzymes: the restriction endonucleases, which are responsible for degrading unmodified foreign DNA, and the modification DNA methyltransferases (methylase or M), which protect endogenous DNA from endonucleolytic digestion by methylating them at the endonuclease recognition sites (7).

Interesting observations have been reported regarding *H. pylori* R-M genes. A novel *H. pylori* gene, iceA (induced when the bacteria contact the host epithelium), was identified recently (8). DNA sequences have revealed two alleles of the iceA locus, iceA1 and iceA2, existing in different *H. pylori* strains. Strains containing iceA1 were found to be significantly associated with peptic ulceration. Increased mucosal concentrations of interleukin-8 were also found in these strains. Surprisingly, iceA1 shares significant sequence similarity with a Type II restriction endonuclease gene, *r.nlaIII* (9).

Second, an interesting phenomenon of phase variation has been linked to the R-M genes in *H. pylori.* Short tandem repeat sequences are subject to loss or gain of a repeat unit, presumably through slipped-strand mispairing during replication. This results in frameshifting, which can alternatively activate or inactivate genes (10). Tetranucleotide repeats were found in a Type III DNA methylase gene and the length of the repeat tract determined the phase variation rate (11). In the case of the *H. pylori* 26695 genome, 27 putative genes that contain simple sequence repeats and that may be subject to phase variation have been identified. These putative phase-variable elements can be divided into three groups: lipopolysaccharide (LPS) biosynthesis, cell-surface-associated proteins and DNA R-M systems (12). For example, the putative Type II R-M system encoded by HP1471–1472 contains a string of 14 G-residues in the HP1471 gene.

Third, *H. pylori* R-M genes are one of the major components of the strain-specific genes. The strain-specific genes are believed to be involved in drug resistance (13) and bacterial surface structure (14), as well as restriction-modification (15). A PCR-based subtractive hybridization method was used to investigate genes that are unique to individual *H. pylori* strains (16, 17). Among the 18 strain-specific genes identified by this method, seven are R-M genes (16). In addition, genome sequence comparison of two *H. pylori* strains showed that R-M system genes account for 15–20% of the strain-specific genes. We reported (18) a biochemical analysis of the Type II R-M systems in *H. pylori* J99. We now report a similar analysis of *H. pylori* 26695 and comparison of these R-M systems with the 16 Type II R-M systems present in strain J99.

Materials and Methods

Bacterial Strains and Growth. *Escherichia coli* DB24 is derived from *E. coli* GM4714 (19), with an additional mutation at the dcm locus introduced via P1 transduction (obtained from E. Raleigh and M. Sibley, New England Biolabs). *E. coli* ER2566-pLysP is an *E. coli* ER2566 derivative that contains a mutant T7 lysozyme gene in the plasmid pACYC184. *E. coli* ER2566-pLysP cells carrying pLT7K with cloned endonuclease genes were grown on LB plates containing 100 μg/ml ampicillin and 20 mM glucose at 42°C overnight. Cells were then removed from the plate and grown in LB media containing 100 μg/ml ampicillin, 34 μg/ml chloramphenicol, and 20 mM glucose at 37°C for 2 h. The cell suspension was equilibrated at 30°C for half an hour before induction with 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG). Following 30 min induction, 200 μg/ml of rifampicin was added to the cell suspension to reduce the background nuclease activity by inhibiting the host *E. coli* RNA polymerase (19). Induced cultures were grown at 30°C for 2 more hours, and then the *E. coli* cells were collected by centrifugation.

Identification of Putative R-M Genes. Putative Type II R-M genes were identified based on the original annotations (4, 5) and our own sequence search (18). First, sequence analysis was used to identify all of the putative methylases in the *H. pylori* 26695 genome. Next,

Abbreviations: R, restriction; M, modification; m6A, N^6-methyladenine; m4C, N^4-methylcytosine; m5C, N^5-methylcytosine.

*To whom reprint requests should be addressed. E-mail: kong@neb.com.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.051612298

the search for additional endonuclease candidates is based on the principle of "guilt by association." Endonuclease and methylase genes in a given R-M system are usually located next to each other. Therefore, an unknown gene with no homolog in GenBank next to a putative methylase gene is a candidate endonuclease gene.

Cloning Potential Type II R-M Systems. DNA sequences coding for all 14 potential Type II R-M systems of *H. pylori* 26695 were obtained from the TIGR web site (http://www.tigr.org/). Those candidate genes were first amplified by using PCR, and then cloned into plasmid vectors. However, the subsequent expression of those restriction endonuclease genes in *E. coli* can be very challenging because the gene products normally are cytotoxic unless the host genomic DNA is completely methylated by the corresponding MTase. To maximize the success rate of expressing these lethal endonucleases, two cloning strategies were used in this study. Putative endonuclease and MTase genes were amplified by PCR using *H. pylori* 26695 genomic DNA and a pair of primers corresponding to the 5′ and 3′ sequences of each pair of R-M genes. The amplified genes were cloned into plasmid pUC19 between the *Bam*HI and *Eco*RI sites. The recombinant plasmid DNA was transformed into *E. coli* ER2683, which lacks endogenous DNA R-M systems (E. Raleigh and M. Sibley, New England Biolabs). These target genes were sequenced to verify that no mutations were introduced during the PCR and cloning processes. Expression of these R-M genes was under the control of a constitutive lac promoter. *E. coli* ER2683 cells carrying the pUC19 vector with both R and M genes were grown in LB broth containing 100 μg/ml ampicillin. Cells were harvested after the OD_{590} reached 0.8–0.9.

The pLT7K system was used to clone potential Type II restriction endonuclease genes directly in the absence of an adjacent gene coding for the cognate MTase (18). Potential restriction endonuclease genes were cloned into pLT7K and sequenced to check for mutations during PCR and cloning. Recombinant plasmids containing the wild-type putative endonuclease genes were then transformed into *E. coli* ER2566 pLysP, which encodes a T7 RNA polymerase. *E. coli* (ER2566pLysP) cells carrying the pLT7K vector with cloned restriction endonuclease genes were grown on LB plates containing 100 μg/ml ampicillin and 20 mM glucose at 42°C overnight. Cells were then removed from the plate and grown in LB media containing 100 μg/ml ampicillin, 34 μg/ml chloramphenicol, and 20 mM glucose at 37°C for 2 h. The cell suspension was equilibrated to 30°C for 30 min before induction with 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG). Following 30 min induction, 200 μg/ml of rifampicin was added to the cell suspension to reduce the background nuclease activity. Two hours later, cells were harvested at 6000 $\times$ *g* for 20 min.

Characterization of Restriction Endonuclease Activity. Restriction endonuclease activity was assayed by using phage lambda and T7 DNA as substrates. The recognition sequences of all active endonucleases were determined by DNA mapping and computer analysis using the programs MAPSORT and GAP (Genetics Computer Group, Madison, WI). The locations of the cleavage sites by the endonuclease were mapped by double digestion of plasmid DNA (pBR322 or pUC19) with known restriction endonucleases and the endonuclease of interest.

Characterization of Methylation Activity. The methylation activity of each potential M gene was determined by a restriction enzyme digestion assay and/or dot blot assay using two rabbit primary antibodies raised against DNA with N^6-methyladenine (m6A) and N^4-methylcytosine (m4C) (18).

Results

We have searched the *H. pylori* DNA sequence looking for the following: a potential DNA MTase (M) gene containing conserved MTase sequence motifs (20, 21) and an adjacent putative restriction endonuclease (R) gene, which either shares sequence similarity with the existing Type II endonuclease or shares no similarity with any known gene in the GenBank. Fourteen potential Type II R-M systems were found by using these criteria and are summarized in Table 1. Type II restriction endonucleases usually do not share sequence similarities with each other (7). Among the 14 putative Type II endonuclease genes, only three (HP91, HP1209, and HP1366) share significant sequence similarities with genes for *Mbo*I, *Nla*III (IceA), and *Mbo*II, respectively (Table 1). In the case of the DNA MTases, they were further annotated into three groups based on the methylation products: m4C, N^5-methylcytosine (m5C), and m6A (7, 20).

The restriction endonuclease activity of each potential R gene was determined by cloning the endonuclease gene and its adjacent methylase gene as a pair into pUC19 vector and then assaying for the presence of DNA cleavage activity in *E. coli* cells expressing the target gene. Sometimes, introducing both endonuclease and methylase genes into a new host cell at the same time could cause problems, because the endogenous DNA may not be sufficiently modified by the incoming MTase if the endonuclease gene is expressed immediately (22). To reduce the possibility of obtaining false-negative results, those putative *H. pylori* endonuclease genes whose gene products showed no detectable endonuclease activity in the pUC19 system were cloned into pLT7K. pLT7K is a cloning/expression vector that features repressor and antisense control elements so that toxic genes can be expressed in *E. coli* under extremely tight control (18).

Table 1. Putative Type II R-M systems in *H. pylori* 26695 strain

HP	Annotation	Putative function	Activity confirmed isoschiz./sequence	Name
49	None	ENase	–	
50	M.*Dpn*IA, 55%, m6A	MTase	+, m6A	M.*Hpy*AVIA
51	M.*Dde*I, 60%, m5C	MTase	+, m5C	M.*Hpy*AVIB
52	None	ENase	–	
53	**None**	**ENase**	**+, ?**	***Hpy*AV**
54	**M.*Hga*I, 55%, m6A**	**MTase**	**+, m6A**	**M.*Hpy*AV**
91	**R.*Mbo*I, 68%**	**ENase**	**+, *Mbo*I**	***Hpy*AIII**
92	**M.*Mbo*I, 75%, m6A**	**MTase**	**+, m6A**	**M.*Hpy*AIII**
262	None	ENase	–	
263	M.*Sca*I, 28%, m4C	MTase	Inactive MTase	
368	None	ENase	–	
369	M.*Cvi*QI, 31%, m6A	MTase	Inactive MTase	
478	M.*Vsp*I, 62%, m6A	MTase	+, m6A	M.*Hpy*AVII
479	None	ENase	–	
481	M.*Fok*I, 49%, m6A	MTase	Inactive MTase	
482	None	ENase	–	
483	M.*Bsp*RI, 64%, m5C	MTase	Inactive MTase	
484	None	ENase	–	
909	None	ENase	–	
910	M.*Hin*cII, 52%, m6A	MTase	+, m6A	
1120	None	ENase	–	
1121	M.*Bsu*FI, 58%, m5C	MTase	+, m5C	M.*Hpy*AVIII
1208	M.*Nla*III, 60%, m6A	MTase	+, m6A	*Hpy*AI
1209	R.*Nla*III, 98%, IceA	ENase	–	
1351	**None**	**ENase**	**+, *Hin*fI**	*Hpy*AIV
1352	**M.*Hin*fI, 78%, m6A**	**MTase**	**+, m6A**	M.*Hpy*AIV
1366	**R.*Mbo*II, 49%**	**ENase**	**+, *Mbo*II**	*Hpy*AII
1367	**M.*Mbo*II, 64%, m6A**	**MTase**	**+, m6A**	M.*Hpy*AIIA
1368	**M.*Mja*I, 41%, m4C**	**MTase**	**+, m4C**	M.*Hpy*AII
1471	*Bcg*I B, 29%,	S	–	
1472	*Bcg*I A, 32%, m6A	R-M	–	

HP, ORF number of *H. pylori* 26695; %, percent of sequence identity; MTase, DNA methylase; ENase, restriction endonuclease; S, specificity subunit. Fully active R-M systems are in bold type. Inactive endonuclease or methylase indicates an ORF with similarity to a known endonuclease or methylase but no function was detected. +, activity detected; –, no activity detected.

MICROBIOLOGY

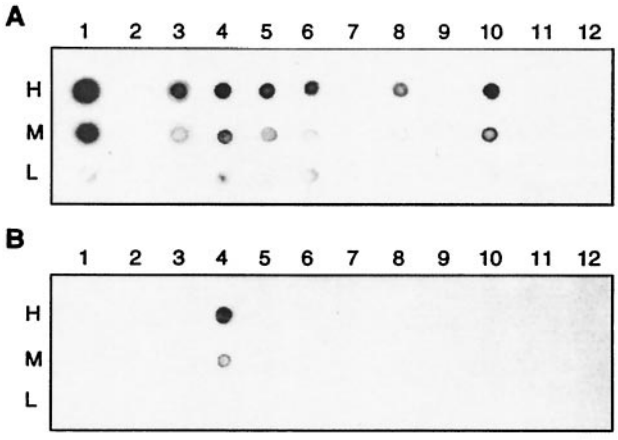

Fig. 1. Antibody assay to measure methyltransferase activities by using rabbit primary antibodies raised specifically against m6A or m4C. Dot blot assays were performed by spotting total DNAs, isolated from clones expressing individual *H. pylori* methylase genes, onto a nitrocellulose BA85 membrane (18). Positive signals were detected by using a secondary anti-rabbit antibody conjugated with horseradish peroxidase (18). Lanes 1–12 represent HP1351–1352; pHKUV5 vector (a negative control); HP91–92; HP1366–1367-1368; HP54; HP50–51; HP263; HP478; HP481; HP910; HP1471–1472; and HP1121, respectively. (*A*) MTase dot blot assay using antibodies against m6A. (*B*) MTase dot blot assay using antibodies against m4C. Three dilutions of DNA samples were spotted: H (high), 0.45 μg; M (medium), 0.15 μg; and L (low), 0.05 μg.

The DNA methylation activity of each potential M gene was determined by a restriction enzyme digestion assay and/or antibody detection assay. The DNA methylase component of a Type II R-M system usually recognizes the same sequence as the corresponding endonuclease. Thus, after a restriction endonuclease is characterized, the methylation activity of its corresponding methylase can easily be determined by digesting the DNA, isolated from cells containing the relevant methylase gene, with the identified endonuclease. Because the majority of the examined *H. pylori* R-M systems showed no detectable endonuclease activity, it was impossible to determine the DNA MTase activity by using the endonuclease digestion method described above. Recently, two polyclonal antibodies raised against m6A or m4C have been developed, and these two antibodies were found to bind specifically to DNA containing these two methylated nucleotides (18).

HP49–50-51. These three ORFs encode two functional MTases and one nonfunctional restriction enzyme-related gene that form part of a Type IIs R-M system. By using antibody analysis we find that the HP50 MTase is an m6A enzyme (Fig. 1*A*, lane 6). The gene sequence is similar to that encoding M.*Mnl*IB (A. Lubys and A. Janulaitis, personal communication), which is a component of a known system that recognizes the sequence 5′-CCTC-3′ (23). The HP50 product, M.*Hpy*AVIA, must methylate the A-residue on the complementary strand of the recognition sequence. Clones expressing M.*Hpy*AVIA are resistant to digestion by *Mnl*I. HP51 is an m5C MTase, M.*Hpy*AVIB, that is also functional. When HP51 is cloned separately into pUC19, it too protects DNA against digestion by *Mnl*I and the gene shows strong sequence similarity to the gene encoding M.*Mnl*IA—the other MTase of the *Mnl*I system. HP52 is almost certainly a remnant of a gene that may have once encoded a functional isoschizomer of *Mnl*I, but in which we can no longer detect activity.

This region in *H. pylori* J99 (JHP43–44) is more complex and shows many rearrangements. The m6A MTase, M.*Hpy*99V, is the product of JHP43 (equivalent to HP50) and is the only functional gene. The adjacent ORF, JHP44, appears to be a chimera between the N terminus of HP52 and the C terminus of HP51. Finally, a new R-M system is inserted after this chimera. This is diagrammed in Fig. 2. It appears as though an earlier strain contained an intact *Mnl*I R-M system and that in J99 it is almost completely rearranged and has lost most of the activities, whereas in 26695, only the R gene appears to have accumulated inactivating mutations (Fig. 2).

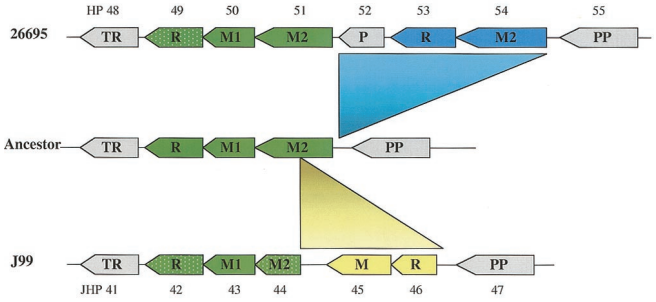

Fig. 2. Schematic diagram showing the possible evolution of HP49–54 in strain 26695 and JHP42–46 in strain J99 from a putative ancestral locus. The shared Type II R-M systems are shown in green. The strain-specific Type II R-M system in strain 26695 is shown in blue, and the strain-specific Type II R-M system in J99 is shown in yellow. Functional genes are shown in solid colors and inactive genes are patterned with white dots. Non-R-M genes are shown in gray. R, restriction endonuclease gene; M, DNA methylase gene; P, putative gene; TR, transcriptional regulator gene; PP, proline permease gene. HP numbers are marked above the genes and JHP numbers are marked below the genes.

HP53–54. These two ORFs encode a fully functional Type II R-M system called *Hpy*AV. HP54 is the gene for an m6A/m5C MTase and has many similar sequences in GenBank. However, it contains an unusually long C terminus of unknown function. The M gene is active because adenine methylation could be detected by m6A antibodies (Fig. 1, lane 5). HP53 has no similar sequences in GenBank, which is typical of an R gene. Restriction endonuclease activity was detectable when the HP53 gene was expressed in *E. coli* cell premodified by the M gene of HP54. The endonuclease activity of *Hpy*AV was purified through several columns. However *Hpy*AV cleaves DNA very frequently and we have not been able to achieve complete digestion on DNA substrates. It probably recognizes a 4- or 5-bp asymmetric sequence. In J99, a totally different R-M system (*Hpy*99II, a *Tsp*45I isoschizomer) is found at this locus (Fig. 2).

HP91–92. HP91 is a fully functional R gene and encodes the restriction enzyme *Hpy*AIII. The gene shows similarities to those of several known isoschizomers, such as *Mbo*I, *Mja*III, and *Dpn*II, all of which recognize GATC. The digestion patterns produced by *Hpy*AIII are identical to those produced by these other enzymes (Fig. 3*A*). The cognate MTase is encoded by HP92 and shows sequence similarities to other known GATC MTases. Antibody tests show that this is an m6A MTase (Fig. 1, lane 3). These genes have a corresponding pair in J99, but there only the M gene is active. The J99 R gene contains multiple frameshifts. However, after correcting the frameshifts the similarity in sequence is very high, which identifies the J99 system as also recognizing GATC.

HP262–263. These two genes do not encode functional proteins, but they do show similarities to known genes. HP263 is most similar to M.*Mja*VI, an m4C MTase that methylates the sequence CCGG. In particular, there is strong similarity within the variable region that encodes sequence specificity. HP262 also shows similarity to the ORF in *Methanococcus jannaschii* that lies next to that encoding M.*Mja*VI, which is probably an ancestral R gene of the CCGG specificity. We conclude that these two genes are inactive versions of an R-M system that once recognized CCGG. Another strain of *H. pylori* (ATCC 49503) encodes an active restriction endonuclease of this same specificity (24).

In J99 this same pair of genes exists (JHP247–248) and the MTase is an active m4C MTase. These observations suggest that the

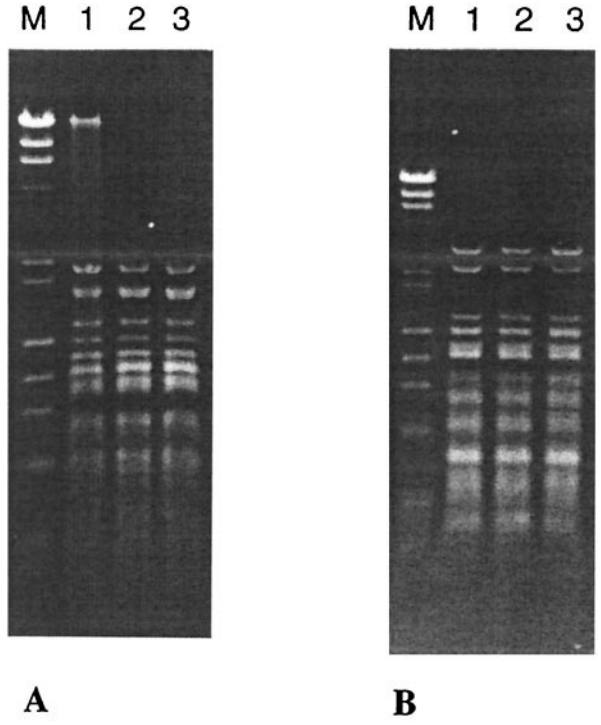

Fig. 3. Determination of the recognition specificities of *H. pylori* endonucleases by double digestion. DNA endonucleotic cleavage activity of isoschizomers of *Mbo*I and *Mbo*II, encoded by HP91–92 and HP1366–1367-1368, respectively. (*A*) Bacteriophage lambda DNA was digested with *Mbo*I (lane 1), *Mbo*I and purified endonuclease *Hpy*AIII (lane 2), or *Hpy*AIII (lane 3). (*B*) Lambda DNA was digested with *Mbo*II (lane 1), *Mbo*II and purified endonuclease *Hpy*AII (lane 2), or *Hpy*AII alone (lane 3).

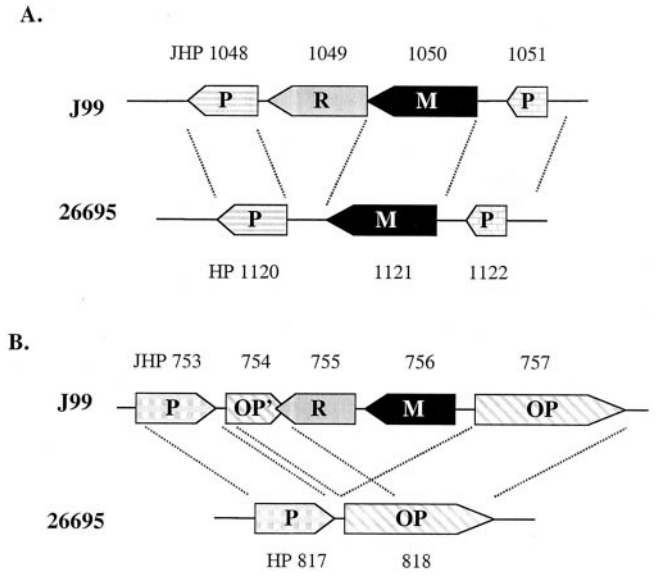

Fig. 4. Schematic diagram showing the alignment of homologous loci. (*A*) The alignment of homologous loci where the strain-specific R-M system genes of JHP755–756 are inserted in strain J99. (*B*) The deletion of an endonuclease gene (homologue of JHP1049) in strain 26695. R, restriction endonuclease gene (shown in gray); M, DNA methylase gene (shown in black). Non-R-M genes are shown in patterned boxes. P, putative gene; OP, osmoprotection protein gene; OP′, 5′ end of a truncated OP gene. JHP numbers are marked above the genes and HP numbers are marked below the genes.

J99 gene, encoding M.*Hpy*99VIII, also has the specificity CCGG. It is unknown which C residue is modified.

HP368–369. These two genes are also inactive. They have homologues in J99, but those genes are in the opposite orientation on the chromosome and at a very different location (JHP1012–1013). All of these genes show strong similarities to the sequences of the *Hpy*188III R-M system, which recognizes the sequence TCNNGA (24). It is likely that these genes in both 26695 and J99 are derived from a previously functional system that also recognized TCNNGA.

HP478–479. By antibody testing, HP478 encodes an active m6A MTase, M.*Hpy*AVII (Fig. 1, lane 8). Its counterpart in J99, JHP430, is inactive and there are 33 predicted amino acid changes within the 545 amino acid coding region. The M genes, HP478 and JHP430, both show over 60% sequence similarity to the gene for *Vsp*I, which recognizes ATTAAT (25). DNA isolated from cells expressing HP478 gene was partially resistant to *Vsp*I endonuclease digestion, suggesting that it is possible that these systems also recognize ATTAAT or a closely related sequence. Because HP477 is likely to encode an outer membrane protein, HP479 is the candidate R gene, although we found no activity when it was cloned into *E. coli*. A similar situation is encountered in J99, where the corresponding, closely related gene, JHP431, is also inactive.

HP481–482. These two genes are inactive, as are their counterparts (JHP433–434) in J99. However, both sets of sequences show strong similarities to other R-M sequences present in GenBank. HP481 and JHP433 are closely related to the gene for M.*Hpy*188I, which recognizes the sequence TCNGA and forms m6A (24) and another system from *H. pylori* strain A17–2 (GenBank accession no. AJ278104) that also recognizes TCNGA (R. Sapranauskas and A. Janulaitis, personal communication). Similarly, the R genes also show similarities and so it is likely that in both 26695 and J99 these genes are derived from a once-functional R-M system recognizing TCNGA.

HP483–484. Both of these genes are inactive, but in J99 the corresponding M gene, JHP435, encodes an active m5C MTase. The corresponding inactive R gene is JHP436. Both M gene sequences show strong similarities to the M gene from the *Hpy*CH4IV R-M system, which recognizes ACGT (24). Thus, we conclude that these genes are another defunct R-M system that once recognized ACGT. Expression of the J99 gene (JHP435) protects against *Hpy*CH4IV confirming its specificity.

HP909–910. This system encodes an active m6A MTase based on antibody tests (Fig. 1, lane 10) and an inactive endonuclease because no activity is found in clones expressing HP909. Both genes show strong similarities to the genes encoding the characterized *Hpy*8I R-M system, which recognizes GTNNAC (A. Lubys, J. Minkute, and A. Janulatis, personal communication; GenBank accession no. AF283672). Thus, this R-M system contains an active M gene and an inactive R gene recognizing GTNNAC. In J99, both genes of the corresponding system JHP845/846 are inactive.

HP1120–1121. HP1121 encodes a functional m5C MTase, M.*Hpy*AVIII, that recognizes the sequence GCGC. In J99 there is a corresponding system, *Hpy*99III (JHP1049–1050), with a fully functional pair of genes. The R gene in J99 is JHP1049, but it has no counterpart in 26695. Instead, the adjacent gene, HP1120, corresponds to JHP1048. Because the flanking genes on the other side of the M gene are conserved between J99 and 26695, we conclude that the R gene from the precursor R-M system has been deleted in 26695 (Fig. 4*A*).

HP1208–1209. HP1208 is similar in sequence to several genes encoding active MTases recognizing CATG, and itself encodes an active m6A MTase, M.*Hpy*AI, recognizing CATG. HP1209 is related to several functional genes encoding active restriction

Table 2. Comparison of activities of Type II R-M systems from two *H. pylori* strains

26696 (HP)			J99 (JHP)	
ORF no.	Name	Sequence	ORF no.	Name
—		GTSAC	***45*-46**	*Hpy*99II
—		CCNNGG	***629*-630**	*Hpy*99IV
—		CGWCG	**755-*756***	*Hpy*99I
1121	*Hpy*AVIII	GCGC	**1049-*1050***	*Hpy*99III
53-*54*	*Hpy*AV	?	—	
91-*92*	*Hpy*AIII	GATC	84-*85*	*Hpy*99VI
1351-*1352*	*Hpy*AIV	GANTC	1270-*1271*	*Hpy*99IX
1366-*1367-1368*	*Hpy*AII	GAAGA	1442	
49-*50-51*	*Hpy*AVI	CCTC	42-*43*-44	*Hpy*99V
262-263		CCGG	247-*248*	*Hpy*99VIII
478-479	*Hpy*AVII	ATTAAT	430-431	
483-484		ACGT	*435*-436	*Hpy*99XI
909-*910*	*Hpy*AIX	GTNNAC	845-846	
1208-1209	*Hpy*AI	CATG	*1131*	*Hpy*99X
368-369		TCNNGA	1012-1013	
481-482		?	433-434	
1471-1472		?	1364-1365	

Active restriction endonuclease genes are shown underlined and active methyltransferase genes are shown in italics. Fully functional R-M systems, which contain both active R and M genes, are highlighted in bold type. ?, unknown recognition sequence.

endonucleases recognizing CATG, but is itself truncated at the N terminus and is inactive when expressed in *E. coli*. In J99, the corresponding M gene is JHP1131 and is active, but there is no corresponding R gene. Instead it is replaced by a locus, termed *ice*A2, found in several different *H. pylori* strains (8).

HP1351–1352. These genes encode a fully active R-M system, *Hpy*AIV, that recognizes the sequence GANTC. The MTase is an m6A MTase. In J99, the corresponding M gene (JHP1271) is active, but the R gene (JHP1270) is not.

HP1366–1368. These genes encode an active R-M system, *Hpy*AII, recognizing the sequence GAAGA. Purified *Hpy*AII cleaves lambda DNA at exactly the same sequence as the *Mbo*II endonuclease (Fig. 3*B*). This R gene shows great similarity to the R gene of the *Mbo*II system (26), and the adjacent gene, HP1367 encoding M.*Hpy*AIIA, closely resembles the m6A MTase of the *Mbo*II system. In *E. coli* cells expressing M.*Hpy*AIIA, overlapping GATC sites become resistant to digestion by *Mbo*I or *Dpn*II, indicating that this MTase methylates the final A residue in its GAAGA recognition sequence. The third gene, HP1368, encodes an m4C MTase that is very similar in sequence to the second MTase of the *Mbo*II system, which has been cloned and sequenced independently (R. Morgan, personal communication). This gene was not reported earlier (26). DNA samples isolated from cells expressing HP1367 and 1368 gave positive reactions with both m6A and m4C antibodies (Fig. 1, lane 4). In J99, the two M genes have both been lost, but an inactive R gene, JHP1442, is found at a different genomic location.

HP1471–1472. The genes of this R-M system show strong similarity to the genes of the *Bcg*I R-M system, an unusual system that contains an M gene fused with an R gene and a separate gene encoding an S subunit of the enzyme responsible for sequence specificity (27). No restriction endonuclease and DNA methylase activities were detected. In the counterpart system in J99 (the completely inactive JHP1364–1365) the fused R-M gene is very similar to the 26695 gene, but the S gene, JH1365, differs in one of the two regions responsible for sequence recognition. We conclude

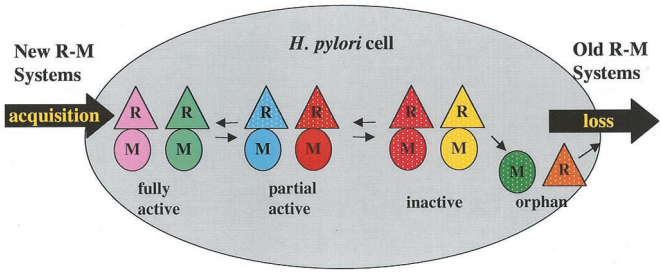

Fig. 5. Schematic diagram showing the dynamic acquisition and loss of R-M system genes. Different R-M systems are shown in different colors. Triangles, R gene; circles, M gene. Orphan genes are not connected to each other. Functional genes are shown in solid colors and inactive genes are patterned with white dots.

that these two *Bcg*I-like systems are derived from a precursor that recognized two different DNA sequences. It should also be noted that both HP1471 and JHP1364 contain a run of 14 G-residues that may be subject to phase variation.

In summary, four of fourteen Type II R-M systems examined were fully functional in *H. pylori*, strain 26695. They displayed both restriction endonuclease activity and methylation modification activity. Four others lacked detectable restriction endonuclease activity, but were positive for methylation activity. Finally, no detectable restriction enzyme activity or methylase activity was observed in the remaining six R-M systems.

Discussion

The biochemical activities of 14 Type II R-M systems from *H. pylori* strain 26695 have been characterized in this study. They are compared with the 16 Type II R-M systems present in the J99 strain of *H. pylori* (18). The results are summarized in Table 2, wherein the systems are broken down into three categories.

The first category shows systems that have both a functional M gene and a functional R gene. There are four such complete systems in each strain, although the recognition specificities of these systems are unique to each strain. The four systems found in J99 consist of three that have genes only found in strain J99, for example JHP755–756 is a unique R-M system, whereas the fourth has an active M gene counterpart in 26695, but no corresponding R gene. In contrast, strain 26695 has one R-M system that has no counterpart in J99 and one system in which the two M genes have been deleted; however, a remnant of an inactive R gene can be seen in J99 and two systems in which the M gene is active in J99, but the R gene is inactive. It is interesting to note that all these strain-specific R-M systems were fully functional. Genomic-sequence comparison of two unrelated *H. pylori* isolates showed that the (G + C) percentage is lower in strain-specific genes (35%) than the reminder of the genome (39%), indicating that they may have been acquired more recently through horizontal gene transfer (5). Analysis of the strain-specific R-M genes in *H. pylori* supports this notion. For example, JHP755–756 is a J99 strain-specific R-M system with a low (G + C) content of 33.9%. Comparison of corresponding loci between J99 and 26695 showed that the JHP755–756 genes were inserted into an osmoprotection protein (OP) gene and a duplication of the 5′ half of the OP gene was observed, which perhaps took place after the insertion event to retain the essential function of the OP gene (Fig. 4*B*).

The second group of R-M systems consists of six systems, each of which contains active methylase genes, but inactive R genes. In one case, the very common M gene encoding an m6A methylase with CATG specificity is active in both strains, but more often a methylase gene active in one strain is inactive in the other. Finally, there are remnants of three systems that are shared by the two strains, but in which neither gene is active. This diversity of R-M systems in *H. pylori* has been documented previously (24), although

earlier the extent of active methylation was not investigated. It should also be noted that the present studies provide only a lower estimate of the total amount of methylation present because, in addition to these Type II R-M systems, there are further potentially active M genes that appear on the basis of sequence analysis to be part of Type I or Type III R-M systems. These are more difficult to assay biochemically. These findings of multiple active M genes without corresponding R genes reinforces previous warnings that restriction cannot be assumed just because site-specific modifications are found (28).

In the case of the Type II systems that have both an R gene and an M gene, but only the M gene is active, we do not presently have data to indicate whether the R gene could be reactivated in a simple fashion. For instance, if the R gene was just one or two mutations away from being active, then these systems might be expected to play a biological role, because within a clonal population active alleles would probably exist. If this were the case, then it would provide an explanation for the maintenance of activity of the M gene because only those members of the population with an active M gene could survive the reactivation of the R gene. The other formal possibility is that the R genes are many mutations away from being fully active and we are observing the slow disappearance of the system. Clearly, loss of the system must first involve the loss of the R gene, because loss of the M gene while the R gene was still active would be a lethal event, and would not be seen. In the latter case, the six systems in Table 2 with active M and inactive R genes will eventually disappear from the strains; unless, of course, the M genes convey some selective advantage outside of their potential role in R-M systems. What that role might be is presently unknown, but given the natural competence of *H. pylori* and the apparently widespread distribution of R-M systems within naturally occurring isolates, perhaps we are observing the selfish nature of R-M systems in action (29). In this scenario, a system that contained an active M gene and an inactive R gene could easily move into another strain and become established, awaiting either the reactivation of the R gene or the appearance of a new, fully active R gene. In this view, strains of *H. pylori* could be viewed as a reservoir for R-M systems, which once established in one strain would spread to other strains and remain established either as fully active systems or in latent form. Certainly, current studies of the distribution of active R-M systems in *H. pylori* would confirm this view, because by directly assaying for active R genes more than 30 different specificities have been discovered, including active examples of all of the identifiable dead genes in strains J99 and 26695 (23).

Some clues that could help decide among these possibilities can be found by examining the nature of the mutations. For instance, in the case of the GATC specificity, the R gene in J99 is inactive as a result of multiple frameshifts. In the case of the GCGC specificity, the R gene in 26695 has been lost completely. For the GANTC specificity, which is fully active in 26695, the R gene in J99 has 20 amino acid differences, but we do not know whether one, two, or more of these are responsible for the loss of activity. Finally, the GAAGA specificity, which is fully active in 26695 has only an inactive remnant of an R gene in J99 and in this case, the inactivation appears to have been caused by a sequence TA-AAAAA, having been converted to a stretch of eight A residues. This is reminiscent of the phenomenon of phase variation (4, 12), although the run of A residues may be a little short to qualify as a *bona fide* possibility. In all of the cases above, it seems more likely that the R genes are in the process of disappearing from the strain rather than being available for reactivation. Thus, the picture that is emerging of *H. pylori* is one in which Type II R-M systems are widespread throughout these strains, suggesting that they are easily acquired because of the natural competence of the strain (30), but that once acquired they are also easily lost through mutation. Therefore, the set of four active systems may reflect the need for a few active systems to overcome the consequence of natural competence, combined with the propensity of R-M systems to act as selfish elements and the natural destruction of these systems by frameshift and point mutations and deletion events. As with other gene systems studies in *H. pylori*, the mutation rates are probably quite high and the whole genome is probably quite fluid (3). On the one hand, *H. pylori* might keep acquiring new R-M systems through horizontal gene transfer. On the other hand, *H. pylori* might lose existing R-M systems by *de novo* mutation and genetic recombination. The current study may provide a snapshot of the dynamic process of acquisition, mutational inactivation and loss of these numerous R-M systems (Fig. 5). Diversified R-M systems in different *H. pylori* strains may be maintained during this process.

We are grateful to Douglas Berg and Martin Blaser for providing *H. pylori* 26695 genomic DNA. We would like to thank A. Janulaitis and A. Lubys for sharing unpublished *Mnl*I results; Shawn Stickel, Lauren Higgins, Nicole Porter, and Michael Dalton for technical help; Ira Schildkraut, Richard Morgan, and Elisabeth Raleigh for helpful discussions; and Donald Comb for his support. Part of this work was supported by National Institutes of Health Grant GM56535 (to R.J.R.).

1. Blaser, M. J. (1996) *Sci. Am.* **274** (2), 104–107.
2. Covacci, A., Telford, J. L., Del Giudice G., Parsonnet, J. & Rappuoli, R. (1999) *Science* **284,** 1328–1333.
3. Wang, G., Humayun, M. Z. & Taylor, D. E. (1999) *Trends Microbiol.* **7,** 488–493.
4. 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., Dougherty, B. A., *et al.* (1997) *Nature (London)* **388,** 539–547.
5. 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., deJonge, B. L., *et al.* (1999) *Nature (London)* **397,** 176–180.
6. Doig, P., de Jonge, B. L., Alm, R. A., Brown, E. D., Uria-Nickelsen, M., Noonan, B., Mills, S. D., Tummino, P., Carmel, G., Guild, B. C., *et al*. (1999) *Microbiol. Mol. Biol. Rev.* **63,** 675–707.
7. Wilson, G. G. & Murray, N. E. (1991) *Annu. Rev. Genet.* **25,** 585–627.
8. Peek, R.M., Jr., Thompson, S. A., Donahue, J. P., Tham, K. T., Atherton, J. C., Blaser, M. J. & Miller, G. G. (1998) *Proc. Assoc. Am. Physicians* **110,** 531–544.
9. Figueiredo, C., Quint, W. G., Sanna, R., Sablon, E., Donahue, J. P., Xu, Q., Miller, G. G., Peek, R.M., Jr., Blaser, M. J. & van Doorn, L. J. (2000) *Gene* **246,** 59–68.
10. Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C. & Moxon, E. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 11121–11125.
11. De Bolle X., Bayliss, C. D., Field, D., van de Ven, T., Saunders, N. J., Hood, D. W. & Moxon, E. R. (2000) *Mol. Microbiol.* **35,** 211–222.
12. Saunders, N. J., Peden, J. F., Hood, D. W. & Moxon, E. R. (1998) *Mol. Microbiol.* **27,** 1091–1098.
13. Davies, J. (1994) *Science* **264,** 375–382.
14. Stroeher, U. H. & Manning, P. A. (1997) *Trends Microbiol.* **5,** 178–180.
15. King, G. & Murray, N. E. (1994) *Trends Microbiol.* **2,** 465–469.
16. Akopyants, N. S., Fradkov, A., Diatchenko, L., Hill, J. E., Siebert, P. D., Lukyanov, S. A., Sverdlov, E. D. & Berg, D. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13108–13113.
17. Kersulyte, D., Mukhopadhyay, A. K., Shirai, M., Nakazawa, T. & Berg, D. E. (2000) *Nucleic Acids Res.* **28,** 3216–3223.
18. Kong, H., Lin, L. F., Porter, N., Stickel, S., Byrd, D., Posfai, J. & Roberts, R. J. (2000) *Nucleic Acids Res.* **28,** 3216–3223.
19. Steen, R., Dahlberg, A. E., Lade, B. N., Studier, F. W. & Dunn, J. J. (1986) *EMBO J.* **5,** 1099–1103.
20. Wilson, G. G. (1992) *Methods Enzymol.* **216,** 259–279.
21. Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V. & Janulaitis, A. (1989) *Nucleic Acids Res.* **7,** 9823–9832.
22. Howard, K. A., Card, C., Benner, J. S., Callahan, H. L., Maunus, R., Silber, K., Wilson, G. & Brooks, J. E. (1986) *Nucleic Acids Res.* **14,** 7939–7951.
23. Roberts, R. J. & Macelis, D. (2000) *Nucleic Acids Res.* **28,** 306–307.
24. Xu, Q., Morgan, R. D., Roberts, R. J. & Blaser M. J. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 9671–9676.
25. Degtyarev, S. Kh., Prikhod'ko, E. A., Prikhod'ko, G. G. & Krasnykh, V. N. (1993) *Nucleic Acids Res.* **21,** 2015.
26. Bocklage, H., Heeger, K. & Muller-Hill, B. (1991) *Nucleic Acids Res.* **19,** 1007–1013.
27. Kong, H. (1998) *J. Mol. Biol.* **279,** 823–832.
28. Ando, T., Xu, Q., Torres, M., Kusugami, K., Israel, D. A. & Blaser, M. J. (2000) *Mol. Microbiol.* **37,** 1052–1065.
29. Kusano, K., Naito, T., Handa, N. & Kobayashi, I. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11095–11099.
30. Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12619–12624.