Movement of DNA sequence recognition domains between non-orthologous proteins

Yoshikazu Furuta[1,2] and Ichizo Kobayashi[1,2,3,*]

[1]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, [2]Institute of Medical Science and [3]Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Minato-ku, Tokyo, 108-8639, Japan

ABSTRACT

Comparisons of proteins show that they evolve through the movement of domains. However, in many cases, the underlying mechanisms remain unclear. Here, we observed the movements of DNA recognition domains between non-orthologous proteins within a prokaryote genome. Restriction–modification (RM) systems, consisting of a sequence-specific DNA methyltransferase and a restriction enzyme, contribute to maintenance/evolution of genomes/epigenomes. RM systems limit horizontal gene transfer but are themselves mobile. We compared Type III RM systems in *Helicobacter pylori* genomes and found that target recognition domain (TRD) sequences are mobile, moving between different orthologous groups that occupy unique chromosomal locations. Sequence comparisons suggested that a likely underlying mechanism is movement through homologous recombination of similar DNA sequences that encode amino acid sequence motifs that are conserved among Type III DNA methyltransferases. Consistent with this movement, incongruence was observed between the phylogenetic trees of TRD regions and other regions in proteins. Horizontal acquisition of diverse TRD sequences was suggested by detection of homologs in other *Helicobacter* species and distantly related bacterial species. One of these RM systems in *H. pylori* was inactivated by insertion of another RM system that likely transferred from an oral bacterium. TRD movement represents a novel route for diversification of DNA-interacting proteins.

INTRODUCTION

Comparisons of proteins indicate that they evolved through movement of domains. However, the elementary steps of these movements have been unclear. In eukaryotes, exon shuffling and alternative splicing generate proteins with switched domains (1). In this work, we report the movement of domain sequences between non-orthologous proteins of a prokaryote species lacking an exon–intron structure. The mobile domains are involved in DNA recognition.

Recognition of DNA sequences by proteins is central to life. Restriction (R) and modification (M) enzymes have provided paradigms for understanding recognition of well-defined DNA sequences (2). In three (I–III) types of restriction–modification (RM) systems (3), an M enzyme methylates DNA at a specific sequence while an R enzyme cleaves DNA lacking this methylation. Type IV restriction enzymes cut DNA that is methylated at a specific sequence (4). DNA methyltransferases brings about three types of modification: m5C, m6A and m4C, and they are mainly grouped into six classes, α to ζ, according to the order of nine conserved motifs and the target recognition domain (TRD) (5,6).

Most of the known Type II systems consist of separate R and M enzymes that independently recognize a target sequence and catalyze reactions (4,7). In M proteins, amino acid sequence motifs common to DNA methyltransferases are well conserved and the TRD is easily identified (6), while R proteins have much less similarity to each other (8).

Type I systems consist of R, M and specificity (S) subunit genes, the products of which form multisubunit enzymes for modification (SM) or restriction (SMR) (9). Sequence recognition is determined by the TRD in the S subunit. The TRD consists of two domains, each of which recognizes half of a bipartite target sequence (10).

Type III systems consist of *res* and *mod* genes. The *mod* gene product alone has M activity, while the complex of the two gene products has R enzyme activity (11). The *mod* subunit is responsible for target recognition and its TRD can be easily identified. Type III *mod* genes in some host-adapted bacterial pathogens are known for diversity in the sequence recognized by the TRD and for

*To whom correspondence should be addressed. Fax: +81 3 5449 5422; Email: ikobaya@ims.u-tokyo.ac.jp

involvement in phase variation of global gene expression (12–14). To date, Type III systems have been annotated by sequence similarity to known Type III enzymes (4), and almost all *mod* genes are classified as β type (REBASE Enzymes, http://rebase.neb.com/cgi-bin/msublist).

RM systems, which limit horizontal transfer of genes, are themselves mobile. Some acquire mobility by traveling with another class of mobile elements such as plasmids and prophages (15–22). Some RM genes are flanked by insertion sequence (IS) elements (23–26). The mobility of other RM systems that are unlinked to a typical mobile element has been suggested by analyses of their phylogenetic relationships, genome contexts and genome comparisons. Phylogenetic trees of RM genes suggest horizontal gene transfer between distantly related prokaryotes (27–29). Genome comparison has revealed insertion of RM systems with long target duplications (30). A large genome inversion was observed next to an RM insertion, suggesting involvement of RM activity in the inversion (31,32). These observations strongly support the nature of RM systems as mobile elements and their contribution to various genome rearrangements. This concept is also supported by experimental analyses (33,34).

The biological significance of RM systems has been mainly explained by their activity as a defense system for host cells against invading DNAs such as those from bacteriophages. Recent work has demonstrated their role beyond defense against invaders, suggesting they are like a watchdog, maintaining epigenetic order. RM systems define specific epigenetic status in a genome by combinatorial methylation of specific genome sequences (35). This epigenetic status regulates the transcriptome (12–14). Alteration of the epigenetic status can lead to cell death by R enzyme activity (36–38). This may help maintain the epigenome and RM systems themselves (39). For example, a host bacterial attack by R enzyme activity might contribute to maintenance of healthy genomes under stressful conditions (38).

Helicobacter pylori are pathogenic epsilonproteobacteria in the human stomach (40) that are known to code for abundant and diverse RM systems (41,42). They are also known for high mutation and homologous recombination rates and for natural competence, which results in high diversity in genomic DNA between isolates (43,44). *Helicobacter pylori* have coevolved since human ancestor started migration from Africa and show wide phylogeographic differentiation (45). By phylogenetic analysis, *H. pylori* are grouped into hspWAfrica, hpEurope, hspAmerind, hspEastAsia and other groups (43). Most earlier studies characterizing their enzymes were restricted to Type II RM systems and to European strains. We compared the complete genome sequences of global *H. pylori* strains for diversity in Type I RM systems and found domain sequence movement (DoMo) within an orthologous gene (46).

In this work, we analyzed diversity in Type III *mod* genes in global *H. pylori* genome sequences. In addition to mobility of the *mod* gene itself, our results revealed various modes of mobility of TRD sequences between *mod* genes.

MATERIALS AND METHODS

Comparison of RM systems

The complete genome sequences used (Table 1) were downloaded from the National Center for Biotechnology Information (NCBI) database as of 1 November 2010, except for those of F16, F30, F32 and F57, which had been obtained by our group (47). The locus tags in the sequences were used as registered in the NCBI database. The genome sequence of strain 908 was not used (except for in core tree construction) because its completeness was not guaranteed.

Sequences of RM systems were downloaded from REBASE (41) as of 15 July 2011. Assignment of a gene as Type III *mod* in *H. pylori* and other *Helicobacter* species was confirmed by significant amino acid sequence similarity by BLASTP (48) (*e*-value $<1e-24$) of at least one allele at a locus with EcoP15 *mod*, which was confirmed experimentally as Type III *mod* (49,50). For further confirmation, a conserved small $PD-(D/E)XK$ motif was also detected in the C-terminal region of Type III *res* genes paired with the *mod* genes or their homologs (51). Nucleotide and amino acid sequences were aligned by mafft (52) and ClustalW (53) with the default parameters. For the gene split at Locus 0 of F57 by conjugative transposon insertion, the nucleotide sequences of HPF57_0278 and HPF57_0312 were concatenated and used for analysis. A homology group of TRDs was defined by clustering with each other with an *e*-value $< 1e-90$ by BLASTN (54).

Phylogenetic tree construction

Homologs of TRDs in the other species were sought by BLASTP (48) against the NCBI nr database, using amino acid sequences of TRD without a stop codon within them as a query. BLASTP hits derived from *H. pylori* were omitted from the results. Hits with an *e*-value $< 1e-50$, which is ∼50% amino acid sequence identity, were retrieved (Supplementary Table S1). The 16S rRNA sequence of the species of the hit was retrieved from the sequence list for the All-Species Living Tree Project (55) and one representative species per genus were chosen for a phylogenetic tree construction (Supplementary Table S2). A phylogenetic tree of the 16S rRNAs was drawn by the maximum likelihood method with Kimura-2 parameter by MEGA (56) with 1000 bootstrap replicates. *Helicobacter felis* was not included because no sequences annotated as 16S rRNA in its genome were full length.

Phylogenetic trees for tree comparison were drawn by the neighbor-joining method with Kimura-2 parameter by MEGA (56) with 1000 bootstrap replicates. The core tree in Figure 5 was redrawn from a published sequence alignment (47) using MEGA (56) with 1000 bootstrap replicates.

Detection of horizontally transferred genes by pentanucleotide word composition

Whole *mod* gene sequences were inferred by pentanucleotide composition whether or not they were horizontally transferred (57). In brief, we first extracted all coding

Table 1. Strains

Strain	Accesion number	Phylogeographic group[a]	Prefix of locus tag
26695	NC_000915	hpEurope	HP
J99	NC_000921	hspWAfrica	jhp
HPAG1	NC_008086	hpEurope	HPAG1_
G27	NC_011333	hpEurope	HPG27_
P12	NC_011498	hpEurope	HPP12_
Shi470	NC_010698	hspAmerind	HPSH_
F16	AP011940	hspEAsia	HPF16_
F30	AP011941	hspEAsia	HPF30_
F32	AP011943	hspEAsia	HPF32_
F57	AP011945	hspEAsia	HPF57_
51	CP000012	hspEAsia	KHP_
52	CP001680	hspEAsia	HPKB_
B38	FM991728	hpEurope	HELPY_
B8	FN598874	hpEurope	HPB8_
Cuz20	CP002076	hspAmerind	HPCU_
PeCan4	NC_014555	hspAmerind	HPPC_
Sat464	CP002071	hspAmerind	HPSAT_
v225d	CP001582	hspAmerind	HPV225_
SJM180	NC_014560	hpEurope	HPSJM_

[a]Based on a phylogenetic tree of the core genes (47).

and non-coding regions from a whole genome sequence and constructed a Markov chain model of coding regions as a training model. Then, the posterior probability that the nucleotide sequence of a gene appeared in the coding regions of the same genome was calculated using the Markov chain model and Bayes theorem. Statistical significance was calculated by the posterior probability of 100 artificial sequences generated based on the Markov chain model. $P < 0.01$ was used as the threshold.

RESULTS

Detecting Type III *mod* genes in *H. pylori* genomes

We retrieved Type III *mod* gene sequences from 19 *H. pylori* complete genomes using REBASE and homology search (see 'Materials and Methods' section). We used only complete genome sequences to ensure finding all possible orthologs and paralogs. *Helicobacter pylori* are known for phylogeographic divergence in their genomes (45). The 19 chosen strains were assigned to hpEurope, hspWAfrica, hspEAsia and hspAmerind groups (Table 1) based on the core phylogenetic tree (confirmed by STRUCTURE analysis, Koji Yahara, personal communication) (47). Each gene had a prefix in the locus tag that was unique to the genome (Table 1). This grouping information was used in the analysis of horizontal transfer of TRD sequences below. The examined *H. pylori* genomes had five orthologous groups of Type III *mod* genes, each at a unique locus (Figure 1; see Supplementary Figure S1 for locus tags).

The *mod* homologs at loci 1 through 4 were found linked to a *res*-like gene with a conserved small PD-(D/E)XK motif at C terminal region. The *mod* homolog at locus 0 is, however, not linked to such a gene in any of these strains when present, although many of its homologs in other species are linked with one. The *mod* homolog at

locus 0 may be a solitary methyltransferase evolutionarily related to Type III RM systems.

Structural variations and insertion of a different RM system likely from an oral bacterium

Type III *mod* loci were found with insertion/deletion of entire genes, truncations by varying length at mononucleotide repeats (14) and nonsense mutations. Some alterations in start and stop codons indicated by small vertical bars in Figure 1 were associated with phase variation (58).

Among the identified loci, locus 4 carried the *mod* gene in only 5 of the 19 strains. Genome context analysis revealed that this absence was due to an apparent substitution by a hypothetical gene in all except a single strain (Figure 2A). The exception was strain PeCan4, in which the *mod* gene was substituted with another Type II RM system and an IS (IS*Hp608*) (Figure 2A). The Type II RM genes had close homology with TdeIII RM genes in *Treponema denticola*, as detected by BLASTP (Figure 2B and C). The amino acid identities in the coding region were 53% for M genes and 62% for R genes. This bacterial species is mainly found in the oral cavity (59), suggesting horizontal gene transfer of this Type II RM system to *H. pylori* from the oral bacterium or a related bacterium. The neighboring IS may have helped this transfer, although we cannot exclude the possibility that the insertion occurred after the transfer.

Diversity in TRD sequences on *mod* genes

Although Locus 0 (previously assigned as mod-1 (60), Figure 1B and Supplementary Figure S2) showed strong conservation in the TRD, the other four orthologous groups at the remaining loci (defined as Loci 1 through 4, previously assigned as mod-3, mod-5, mod-4, mod-2, respectively (60), Figure 1C–F and Supplementary Figures S3–S6) had significant sequence diversity in their TRDs. Allelic variation at locus 2 (Figure 1D) has been reported (14,61). In this work, we defined TRD homology groups using clustering of TRD sequences after BLASTN analysis. TRD sequences were clustered in the same homology group when the *e*-value in BLASTN was $< 1e - 90$. We identified 22 distinct TRD homology groups in all, with two to eight distinct TRD homology groups at each of the four loci, among the 19 strains.

Movement of TRD sequences between *mod* genes of different homology groups at different loci

The diversity of TRD sequences within the same orthologous group (and at the same locus) (Figure 1) can be explained by allelic homologous recombination at the conserved regions flanking the TRD sequences, referred to as non-TRD regions. In addition, we found that some TRDs were shared by *mod* genes of different orthologous groups at different loci (Figure 1). TRD homology groups A and C were found at loci 1 and 3, while TRD homology group D was found at loci 1, 2 and 4. This suggested movement of a TRD sequence between different orthologous groups at different loci.
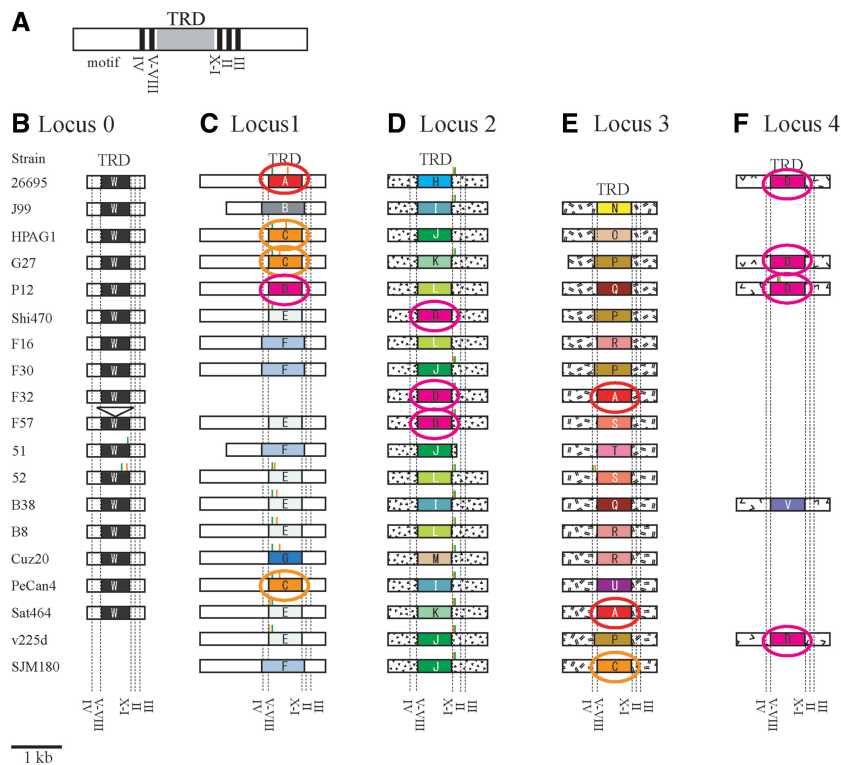
Figure 1. TRDs of Type III *mod* genes in *H. pylori* complete genomes. (**A**) Relative positions of conserved motifs and TRD in Type III *mod* genes. Roman numerals indicate the conserved motifs of DNA methyltransferases (6). (**B**) Locus 0. HP0260 homologs. A large triangle on the homolog in strain F57 represents insertion of a conjugative transposon. (**C**) Locus 1. HP1369 homologs. (**D**) Locus 2. HP1522 homologs. (**E**) Locus 3. jhp1296 homologs. (**F**) Locus 4. HP0593 homologs. Members of the same TRD homology group are in the same color. Small vertical bar in orange, start codon; small vertical bar in green, stop codon generated by a frameshift mutation. For locus tags, see Supplementary Figure S1.

We hypothesized about the mechanism underlying such movements. Homologous recombination involving most of the non-TRD regions cannot be used for movement because the non-TRD regions were not conserved between different orthologous groups. By detailed comparison of the flanking sequences, we found that the movement could be explained by recombination using a short DNA sequence similarity at the regions flanking the TRD sequences. These are common among the methyltransferases of different orthologous groups (Figure 3).
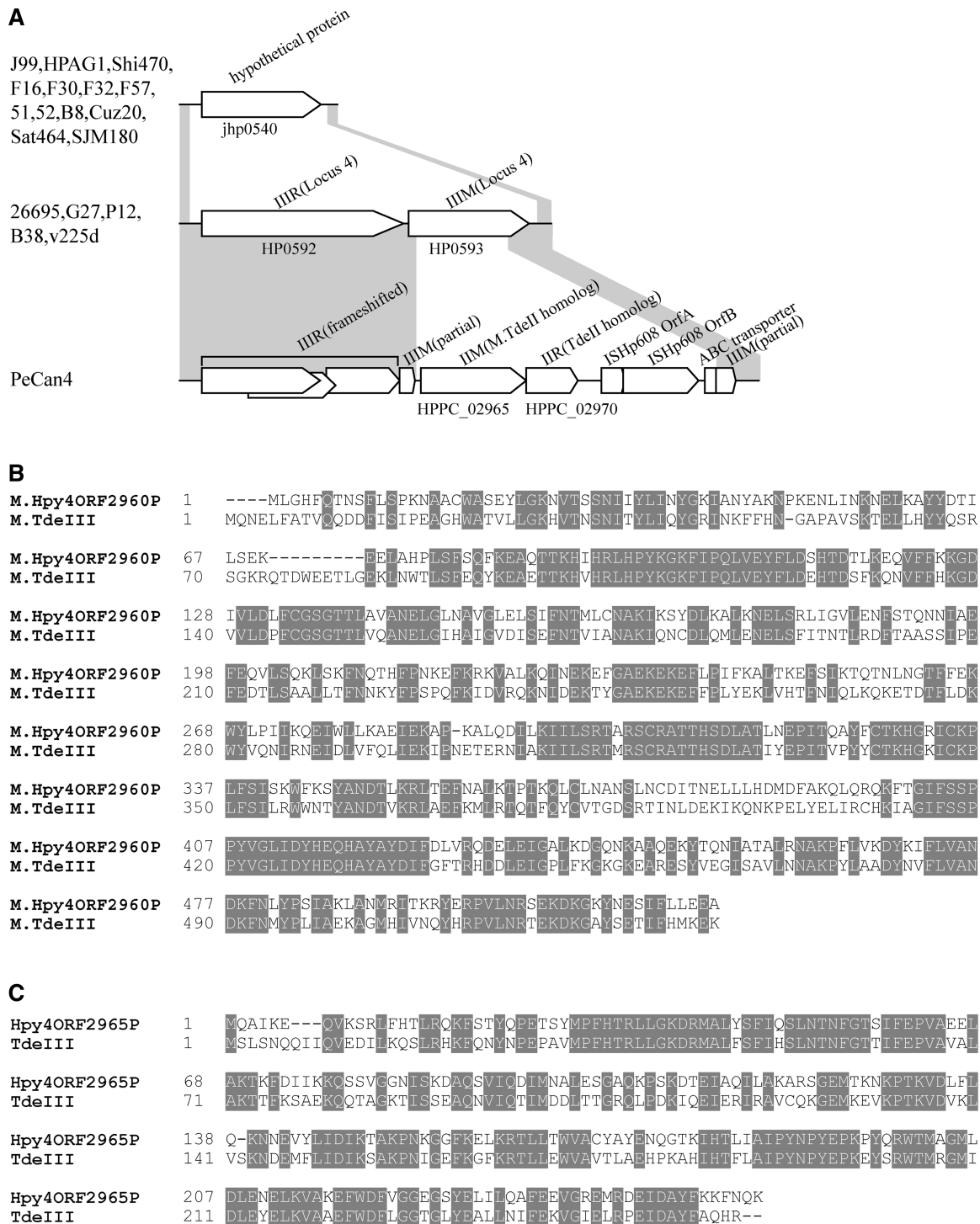
For example, the movement of group A between loci 1 and 3 could be explained by the sequence similarity of 13 bp at the 5′-side of the TRD sequences, and 117 bp at the 3′-side (Figure 3A). The movement of group C between loci 1 and 3 could be also explained by 17 bp of sequence similarity at the 5′-side and 117 bp at the 3′-side of the TRD sequences (Figure 3B). The cases for Group A and Group C are similar for the sequences involved.

For group D movement between loci 2 and 4, the sequence similarity was 32 bp at the 5′-side and 57 bp at the 3′-side of the TRD sequences (Figure 3C). In contrast, the sequence similarity between loci 1 and 2 was 8 bp at the 5′-side and 26 bp at the 3′-side (Figure 3D). For the 5′-boundary, the expansion of a poly-G repeat at locus 1 might have disrupted a longer region of similarity. Poly G repeats were associated with two additional recombination areas (Figure 3A and B). We do not know whether these were related to the process or a consequence of the recombination.

Phylogenetic incongruence between TRD and non-TRD regions consistent with TRD movement

Movement of a TRD sequence between different loci (orthologous groups) might lead to phylogenetic incongruence between the TRD and the remaining non-TRD gene regions. Phylogenetic trees were compared for all four loci (Figure 4; see Supplementary Figure S7 for locus tags and bootstrap values). TRD homology groups A, C and D, each of which formed a cluster in the TRD tree, were connected with more than one cluster of the non-TRD regions' tree, consistent with the movement of these TRD sequences between orthologous groups (loci). In addition, clustering in the TRD tree seemed to be mostly independent of locus. This was in contrast to clustering in TRD homology group W and non-TRD homology group at Locus 0, where all members were clustered into a single group. These results are consistent with the movement of TRD sequences between loci (homology groups with respect to non-TRD regions) during evolution. In other words, the diversity of TRD appeared to have occurred through acquiring a new sequence rather than only through accumulation of mutations in an ancestral TRD sequence unique to each locus (orthologous group).

A

B

```
M.Hpy4ORF2960P    1   ----MLGHFQTNSFLSPKNAACWASEYLGKNVTSSNIIYLINYCKIANYAKNPKENLINKNELKAYYDTI
M.TdeIII          1   MQNELFATVQQDDFISIPEAGHWATVLLGKHVTNSNITYLIQYGRINKFFHN-GAPAVSKTELLHYYQSR

M.Hpy4ORF2960P    67  LSEK---------EELAHPLSFSQFKEAQTTKHIHRLHPYKGKFIPQLVEYFLDSHTDTLKEQVFFKKGD
M.TdeIII          70  SGKRQTDWEETLGEKLNWTLSFEQYKEAETTKHVHRLHPYKGKFIPQLVEYFLDEHTDSFKQNVFFHKGD

M.Hpy4ORF2960P    128 IVLDLFCGSGTTLAVANELGLNAVGLELSIFNTMLCNAKIKSYDLKALKNELSRLIGVLENFSTQNNIAE
M.TdeIII          140 VVLDPFCGSGTTLVQANELGIHAIGVDISEFNTVIANAKIQNCDLQMLENELSFITNTLRDFTAASSIPE

M.Hpy4ORF2960P    198 FEQVLSQKLSKFNQTHFPNKEFKRKVALKQINEKEFGAEKEKEFLEIFKALTKEFSIKTQTNLNGTFFEK
M.TdeIII          210 FEDTLSAALLTFNNKYFPSPQFKIDVRQKNIDEKTYGAEKEKEHFPLYEKLVHTENIQLKQKETDTELDK

M.Hpy4ORF2960P    268 WYLPIIKQEIWLLKAEIEKAE-KALQDILKIILSRTARSCRATTHSDLATLNEPITQAYFCTKHGRICKP
M.TdeIII          280 WYVQNIRNEIDLVFQLIEKIENETERNIAKIILSRTMRSCRATTHSDLATIYEPITVPYYCTKHCKICKP

M.Hpy4ORF2960P    337 LFSISKWFKSYANDTLKRLTEENALKTPTKQLGLNANSLNCDITNELLLHDMDFAKQLQRQKFTGIFSSP
M.TdeIII          350 LFSILRWWNTYANDTVKRLAEEKMLRTQTFQYCVTGDSRTINLDEKIKQNKPELYELIRCHKIAGIFSSP

M.Hpy4ORF2960P    407 PYVGLIDYHEQHAYAYDIFDLVRQDELEIGALKDGQNKAAQEKYTQNIATALRNAKFFLVKDYKIFLVAN
M.TdeIII          420 PYVGLIDYHEQHAYAYDIFGFTRHDDLEIGPLFKCKGKEARESYVEGISAVLNNAKPYLAADYNVFLVAN

M.Hpy4ORF2960P    477 DKFNLYPSIAKLANMRITKRYERPVLNRSEKDKGKYNESIFLLEEA
M.TdeIII          490 DKFNMYPLIAEKAGMHIVNQYHRPVLNRTEKDKGAYSETIFHMKEK
```

C

```
Hpy4ORF2965P    1   MQAIKE---QVKSRLFHTLRQKFSTYQPETSYMPFHTRLLGKDRMALYSFIQSLNTNFGTSIFEPVAEEL
TdeIII          1   MSLSNQQIIQVEDILKQSLRHKFQNYNPEPAVMPFHTRLLGKDRMALFSFIHSLNTNFGTTIFEPVAVAL

Hpy4ORF2965P    68  AKTKFDIIKKQSSVGGNISKDAQSVIQDIMNALESGAQKPSKDTEIAQILAKARSGEMTKNKPTKVDLFL
TdeIII          71  AKTTFKSAEKQQTAGKTISSEAQNVIQTIMDDLTTGRQLPDKIQEIERIRAVCQKGEMKEVKPTKVDVKL

Hpy4ORF2965P    138 Q-KNNEVYLIDIKTAKPNKGGFKELKRTLLTWVACYAYENQGTKIHTLIAIPYNPYEPKPYQRWTMAGML
TdeIII          141 VSKNDEMFLIDIKSAKPNIGEFKGFKRTLIEWVAVTLAEHPKAHIHTFLAIPYNPYEPKEYSRWTMRGMI

Hpy4ORF2965P    207 DLENELKVAKEFWDFVGGEGSYELILQAFEEVGREMRDEIDAYEKKFNQK
TdeIII          211 DLEYELKVAAEFWDFLGGTGLYEALLNIFFKVGIELRPEIDAYEAQHR--
```

Figure 2. Evolution at Locus 4. (**A**) Structure in various strains. IIM, Type II system M gene; IIR, Type II system R gene; IIIM, Type III system *mod* gene; IIIR, Type III system *res* gene. (**B**) Alignment of amino acid sequences between HPPC_02695 (Type II M) and M. TdeIII derived from *T. denticola*. (**C**) Alignment of amino acid sequences between HPPC_02670 (Type II R) and TdeIII derived from *T. denticola*. Identical residues in alignments are shaded.

Movement of *mod* genes between *H. pylori* genomes inferred from phylogenetic incongruence

To follow any mobility of the *mod* genes with respect to the entire genome, the phylogenetic trees of the non-TRD regions at each locus were compared with a phylogenetic tree of concatenated core genes, which in some sense reflects the overall evolution of the entire genome (Figure 5). Some groups showed entanglement compared with the core tree.

For Locus 1 (Figure 5B), the major clustering into two in the non-TRD regions was different from the clustering
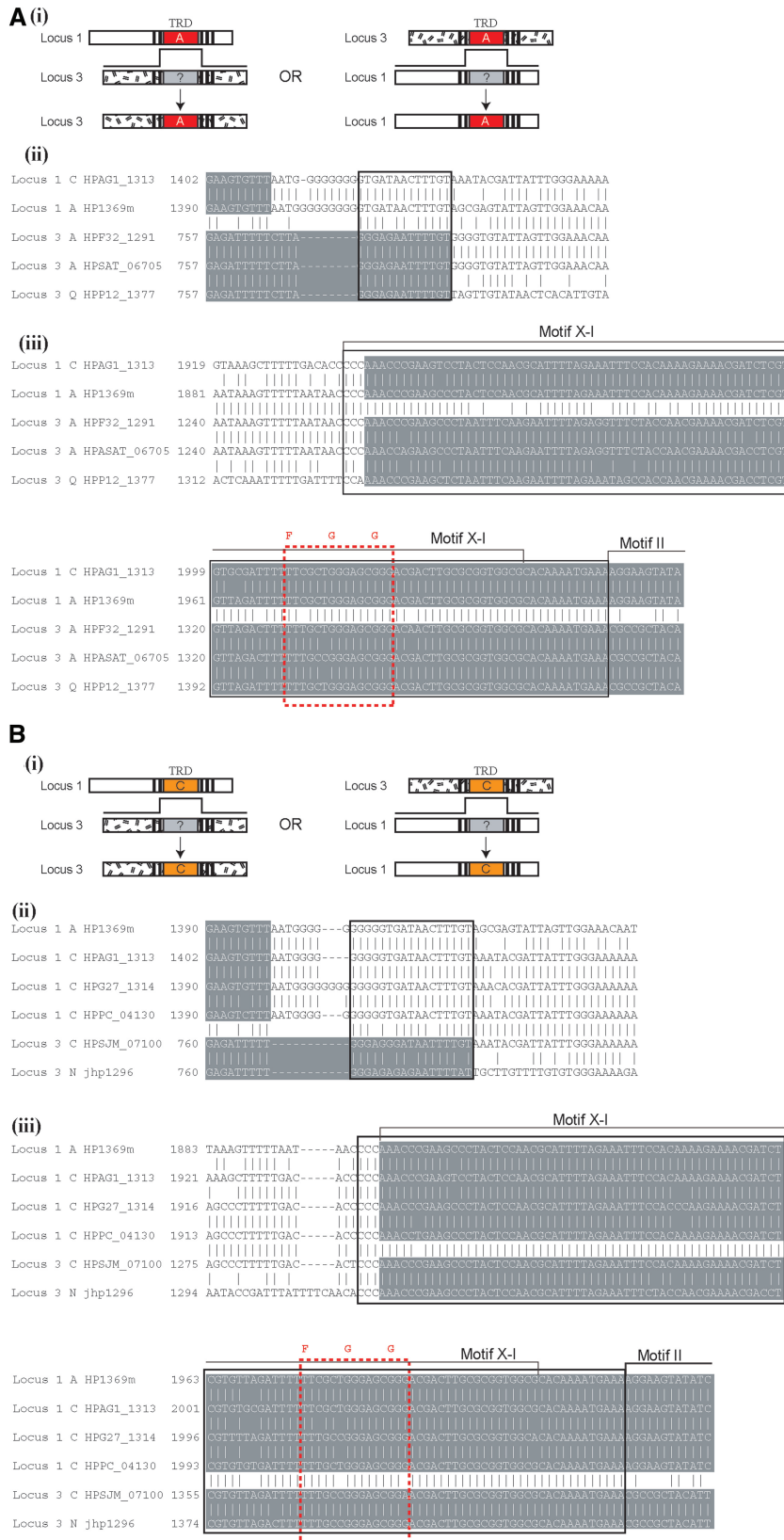
Figure 3. Sequence alignments at suggested recombination sites for TRD replacement. (**A**) (i) Recombination scheme, (ii) 5′-side alignment and (iii) 3′-side alignment of group A. (**B**) (i) Recombination scheme, (ii) 5′-side alignment and (iii) 3′-side alignment of group C. (**C**) (i) Recombination scheme, (ii) 5′-side alignment and (iii) 3′-side alignment of group D at loci 2 and 4. (**D**) (i) Recombination scheme, (ii) 5′-side alignment and (iii) 3′-side alignment of group D at loci 1 and 2. Gray, conserved sequences at each locus; boxed, sequences for recombination; red hatched box, sequence corresponding to conserved amino acid sequence of motif I, FxGxG.

(continued)
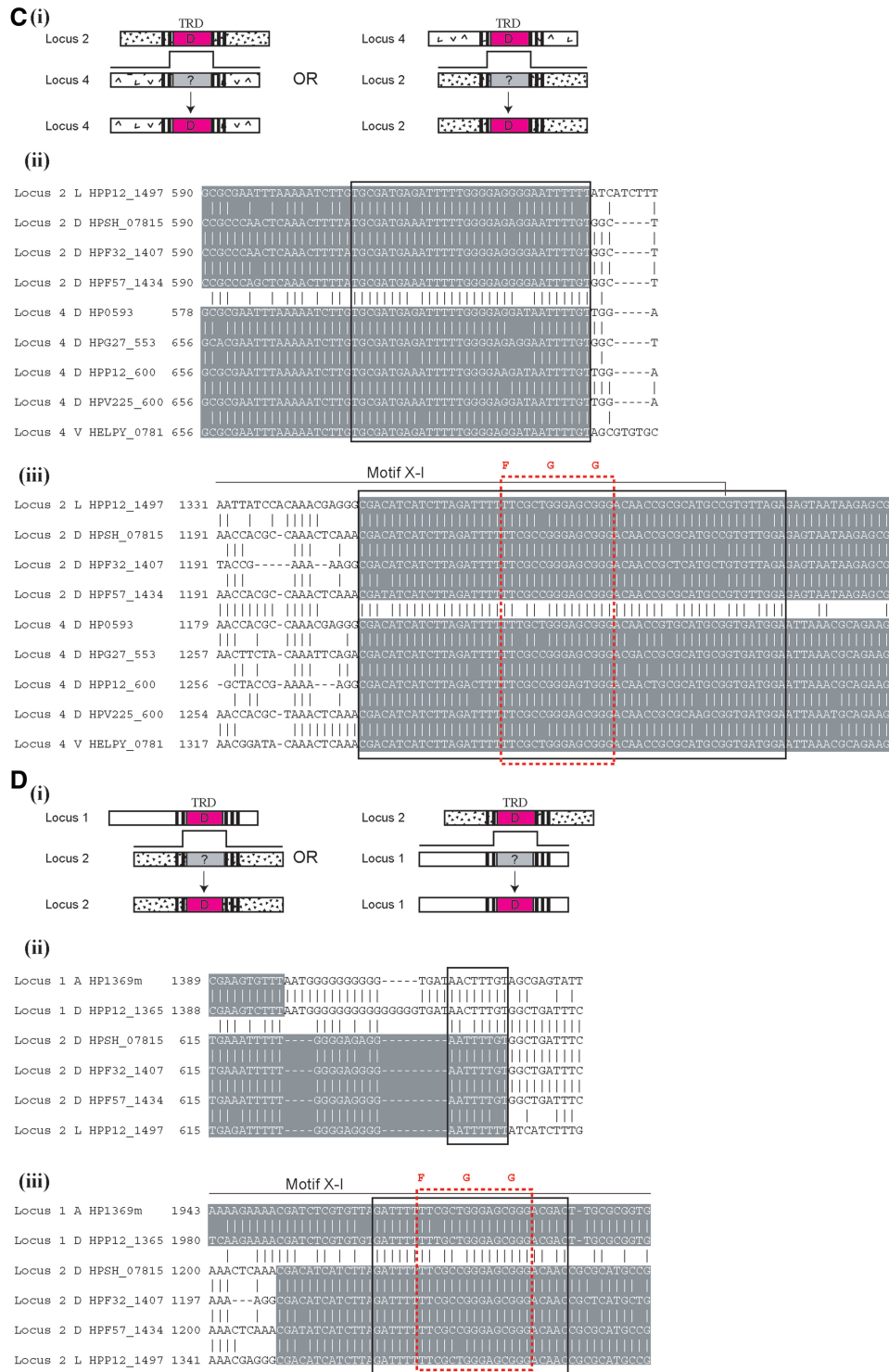
Figure 3. Continued.

in the core tree. Separation of SJM180 from the other hpEurope strains and clustering with several hspEastAsian strains might reflect a horizontal transfer from an hspEastAsia strain to an hpEurope strain.

For Locus 2 (Figure 5C), PeCan4 was separated from the hspAmerind cluster and included in the hpEurope cluster. A transfer might have occurred from an hpEurope strain to a PeCan4 ancestor.

Tree comparison of Locus 3 (Figure 5D) showed a more complex pattern, which separated all the major phylogeographic groups (hpEurope, hspAmerind, hspEastAsia) into more than one of the three major clusters.
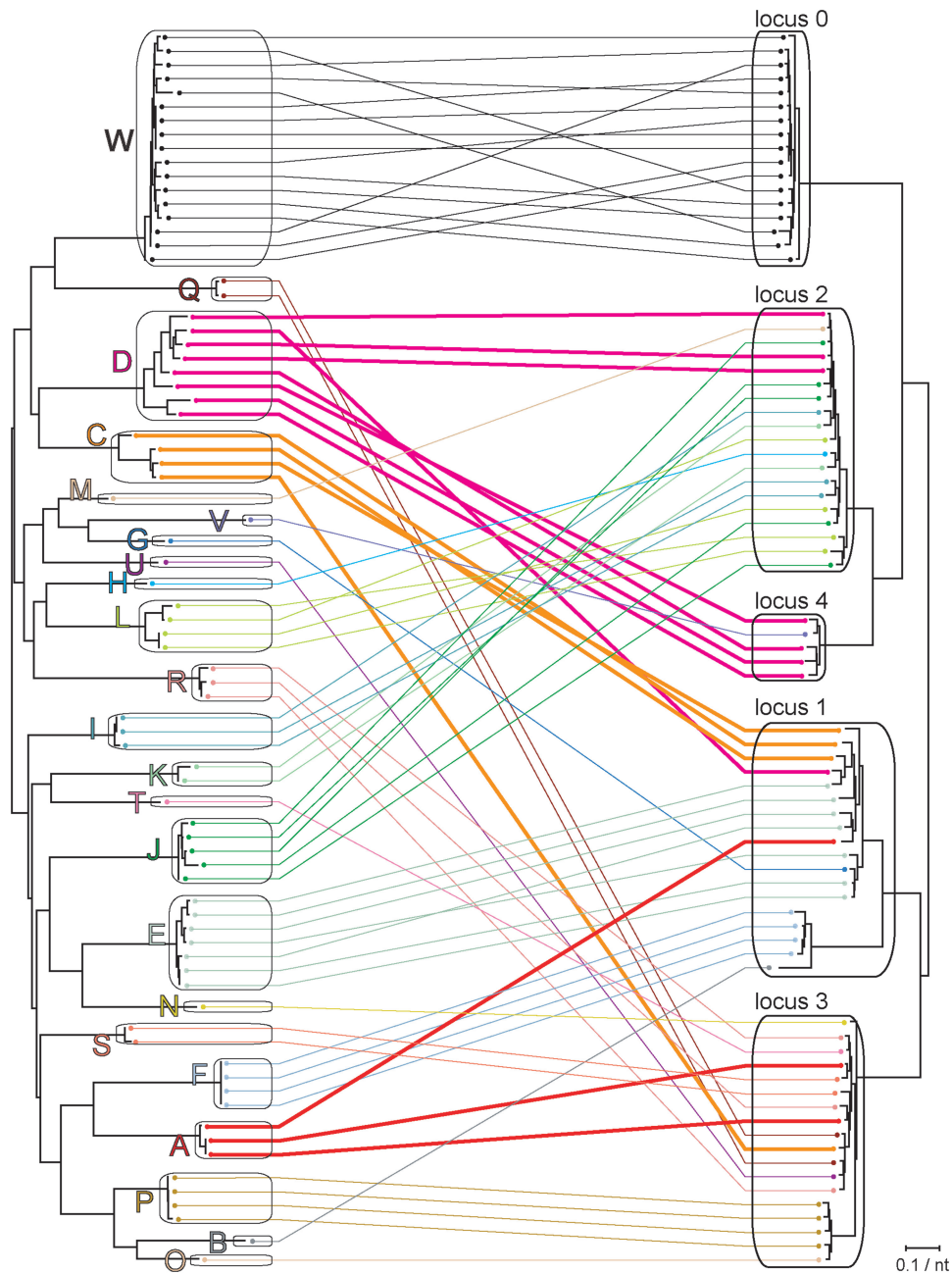
Figure 4. Comparison of phylogenetic trees of the TRD and non-TRD regions of Type III *mod* genes at all loci. Left, TRD trees; right, non-TRD region. Line colors correspond to TRD colors in Figure 1. TRD groups shared by different loci are connected with thicker lines. For locus tags and bootstrap values, see Supplementary Figure S7.

These apparent incongruities between the trees of the non-TRD regions and the core genome suggested recombination of the *mod* genes with the remainder of the genome, consistent with the high degree of mutual homologous recombination in *H. pylori* (43). Combined with TRD mobility between *mod* genes, the mobility of *mod* genes revealed multilevel mobility in RM systems.

From these limited results, we have not so far obtained evidence to relate the two levels of mobility: TRD movement between *mod* genes and *mod* movement between genomes.

We do not know whether *mod* gene moved by itself or together with linked genes. A hypothesis of transfer of many genes other than the *mod* gene at locus 2 from the European/African lineages to PeCan4 would be consistent with the position of this strain in the core tree.

mod genes and TRD sequences in other *Helicobacter* species

We investigated the origin of the extremely diverse TRD sequences of *mod* genes in *H. pylori*. Rather than simple accumulation of mutations and intraspecific transfer, we

suspected acquisition by horizontal gene transfer from other bacterial species as suggested for the Type I specificity subunit (62).

First, we searched for homologs of *mod* genes in *Helicobacter* species other than *H. pylori* to determine history of gain/loss of *mod* genes themselves (Figure 6

and Supplementary Figure S8). The number of *mod* genes per genome varied from 0 to 5. More than half showed homology to the non-TRD region of *H. pylori* *mod* genes. These *mod* homologs carry a TRD sequence from the various TRD homology groups found above in *H. pylori*. In addition, 10 novel groups for TRD and five
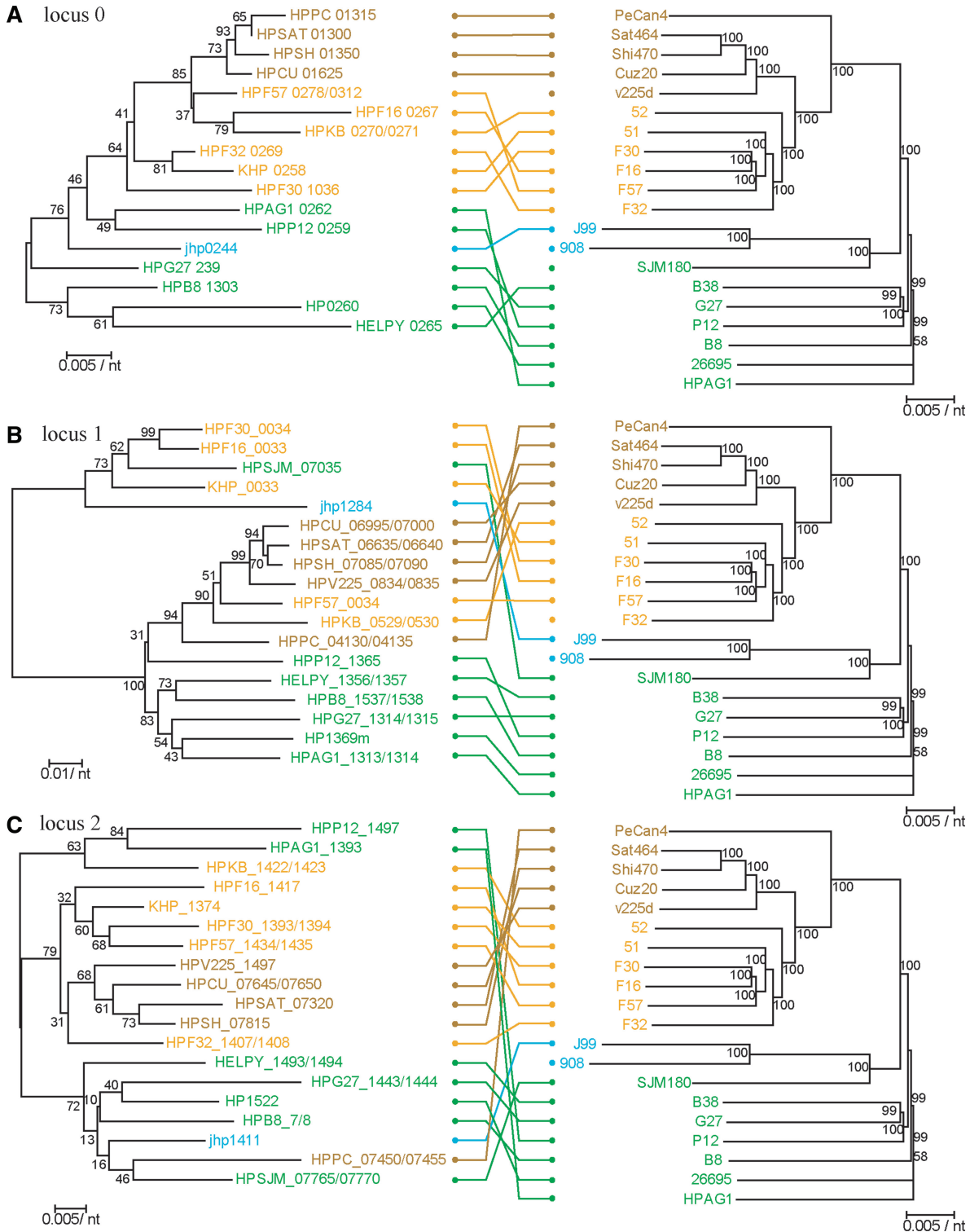
Figure 5. Comparison of phylogenetic trees of non-TRD regions of *mod* genes at each locus and the core genome. Left, non-TRD regions; right, core genome. Numbers indicate bootstrap values. Colors indicate phylogenetic groups in Table 1.

(continued)
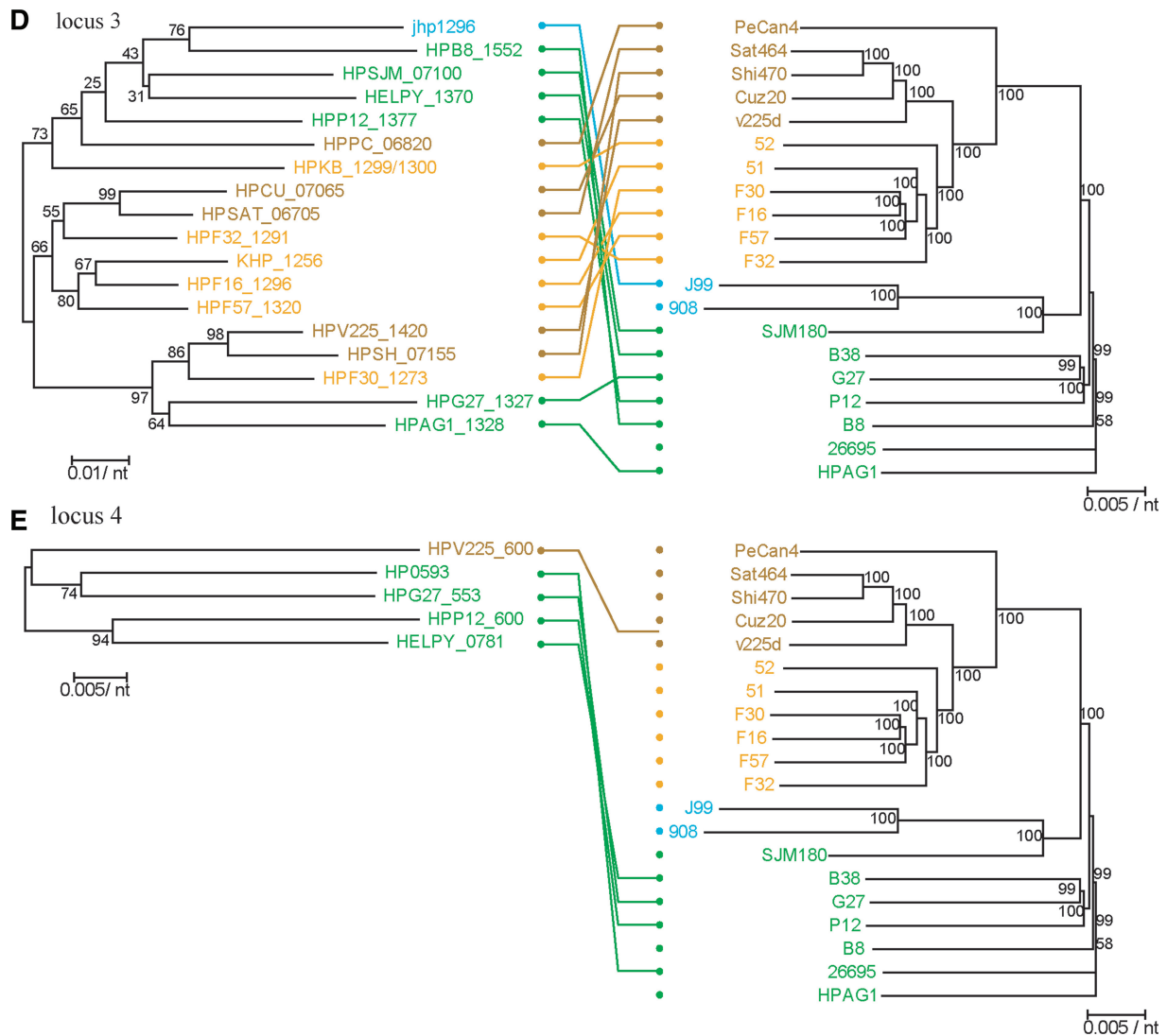
Figure 5. Continued.

novel groups for non-TRD were found among the non-pylori *Helicobacter* species. To determine the gain/loss history, we listed these *mod* homologs in a phylogenetic tree of *Helicobacter* species (Figure 6).

Homologs of *mod* at *H. pylori* Locus 0 and *mod* at locus 4 were observed only in *H. pylori* and *H. acinonychis*. The simplest explanation for this pattern is that these two subfamilies were acquired after separation of *H. cetorum* from the common ancestor of *H. pylori* and *H. acinonychis* and before separation of the latter two species. Because homologs of Locus 2 *mod* were also observed in *H. cetorum*, those homologs might have been acquired before separation of *H. cetorum* and the other two species.

Homologs of Locus 3 *mod* gene were found in *H. pylori*, *H. acinonychis*, *H. cetorum* and *H. cinaedi*, a species distantly-related to *H. pylori* but that infects humans. Horizontal transfer might have occurred between *H. cinaedi* and the common ancestor of *H. pylori* and *H. cetorum*.

Five *mod* homology groups (designated groups 5–9) were not homologous to any non-TRD region of the *H. pylori mod* genes identified (Figure 6). They were found in *H. acinonychis*, *H. cetorum*, *H. suis*, *H. canadensis*, *H. bilis* and *H. cinaedi*. Species-specific *mod* homology groups were found even within closely-related species such as *H. pylori*, *H. acinonychis* and *H. cetorum*, which suggested frequent gain/loss events of whole *mod* genes.

A phylogentic tree of the newly identified 10 TRDs and *H. pylori* TRDs (Supplementary Figure S8) revealed their extensive diversity. None of the new TRDs was closely related to another. The tree as a whole was consistent with extensive horizontal transfer of TRDs between *Helicobacter* species. Interestingly, TRD homology group D with groups (loci) 1, 2 and 4 *mod* genes in *H. pylori* was found to be associated with group 9 *mod* in *H. suis*. This observation suggested that TRD movement between non-orthologous *mod* genes could occur not only in *H. pylori* but also in other species and contribute to distant horizontal transfer.
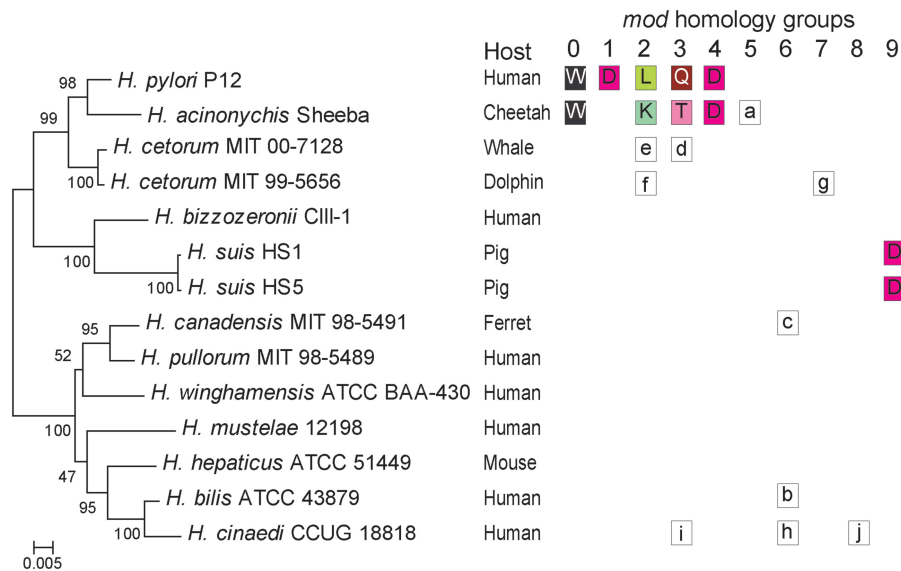
Figure 6. Distribution of TRD homologs in *Helicobacter* species. Each column in the right represents a homology group found in the *Helicobacter* species. Groups 0 through 4 correspond to *mod* genes at Loci 0 through 4 in *H. pylori*. Colors in a square represent the TRD sequences as in Figure 1.

TRD sequences in distantly related bacteria

Next, to examine the possibility of the transfer of TRD sequences from/to distantly related species, we searched sequences that showed similarity to the diverse TRD sequences by BLASTP analysis against the nr database (Figure 7). For some homology groups of TRD, similar sequences were detected in species distant from *H. pylori*, but not from epsilonproteobacteria such as other *Helicobacter* and *Campylobacter* species. This suggested distant horizontal transfer as opposed to vertical transfer. In particular, TRD homology group R had strong sequence similarity even at the nucleotide level to a gene in *Haemophilus* (around *e*-value $1e - 80$ by BLASTN) and group U to one in *Mycoplasma* (*e*-value $2e - 60$ by BLASTN). These results strongly supported a relatively recent horizontal transfer of TRD sequences between these pairs of distantly related species.

We also determined horizontal gene transfer based on pentanucleotide word compositions of entire open reading frames (see 'Materials and Methods' section). This method detects coding sequences recently transferred from a distantly related organism. All *mod* genes in *H. pylori* were judged as recently transferred except for TRD C and TRD I. The TRD group C was found in many bacterial groups, so we could not determine which served as the donor and which as the recipient. TRD group O might have moved from *Fusobacterium* to *H. pylori* and *Neisseria*, while TRD group T might have moved from Ureaplasma to epsilonproteobacteria.

DISCUSSION

### Movement of TRD between genes for Type III RM systems

We analyzed Type III *mod* genes at the nucleotide sequence level in global *H. pylori* genomes. We found

mobility of the TRD sequences between different *mod* orthologs at different loci. The mechanism underlying the TRD movement was suggested to be recombination at 8–117 bp of similar sequences flanking TRD regions. The *mod* genes analyzed in this work all belonged to the β group, as do most of the other *mod* genes, based on the arrangement of the methyltransferase motifs and TRD (49). The TRD was flanked by motif IV–VIII at the 5′-side and X–III motif at the 3′-side. Recombination apparently took advantage of the conservation of DNA sequences at both the flanking regions that encode the conserved amino acid motifs.

A related mode of TRD diversification by recombination was previously observed in the specificity subunit of Type I RM systems and other genes (46,63–65). Some Type I specificity subunits consist of two TRDs flanked by the same pair of 19–49 bp sequences. Taking advantage of these repeat sequences for recombination, a TRD sequence can replace another TRD at the same or at another TRD site. The target TRD site can be at a separate locus. The movement between the two TRD sites is named DoMo (46). TRD sequence movement by DoMo in the Type I specificity subunit is restricted to the same orthologous group, whether between the same locus or between different loci. However, the TRD movement of Type III *mod* we found here is unique because it takes place between different orthologous groups, taking advantage of the weak homology at the motif sequences conserved among DNA methyltransferases.

Comparison with related reactions

Movement of sequences between homologous genes at different loci is known as gene conversion (66). It is frequently observed in *H. pylori* outer membrane protein genes such as those in *bab* and *sab* families (67,68). Gene conversion from an unexpressed to an expressed locus
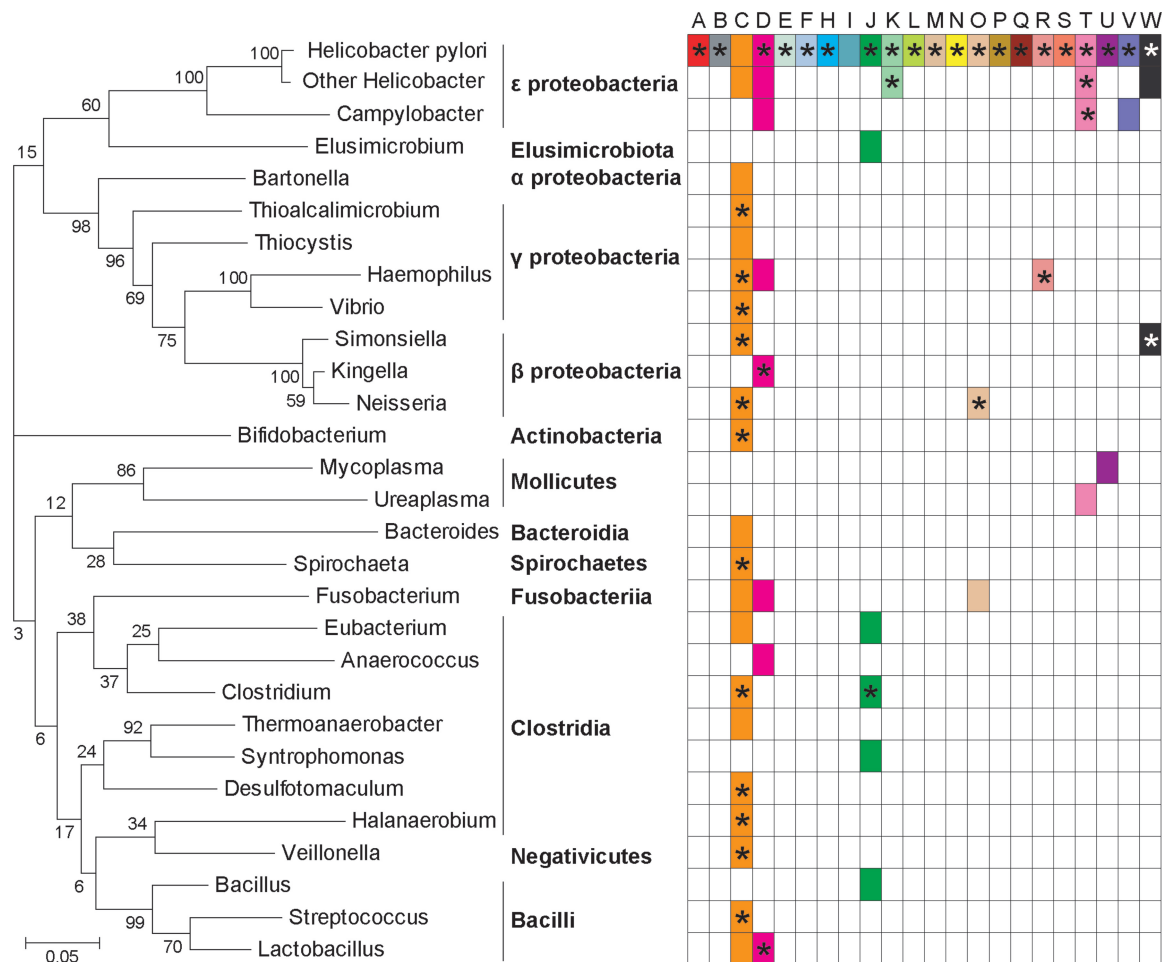
Figure 7. TRD homologs detected in distant species. Left, a 16S rRNA-based phylogenetic tree of a genus where a TRD homolog was detected. Middle, classes of these genera. Right, presence (color) or absence (white) of a homolog of each TRD group in each genus. Note that they are colored when at least one of the species in the genus was detected with similar sequence to TRD groups. Asterisk: at least one of similar genes in a genus was detected as horizontally transferred by pentanucleotide composition analysis of its entire open reading frame.

mediates antigenic variation of outer membrane proteins and pili in several bacteria (69–71).

The TRD movement reported in this work is unique in the low sequence similarity of the recombining regions (13–52% nucleotide sequence identity) between different orthologous groups, which encode conserved amino acid motifs of DNA methyltransferases. Most of other examples of gene conversion use long flanking regions conserved between genes as recombination sites. The movements use relatively long similar sequences at 3'-side of TRD, but short (13–32 bp) similar sequences at the 5'-side. This might explain the apparently lower frequency of TRD movement observed between different orthologous groups.

An example similar to the short conserved-motif-driven gene conversion was described for rearrangement in a tandem paralog cluster in *Staphylococcus aureus* (72). In this case, conserved motif sequences in paralogs in tandem are used as sites for unequal homologous recombination. Similar unequal recombination was not found for Type III RM systems here but has been observed for Type I RM systems (46). Another example of domain shuffling in an intronless gene was reported in albumin-binding genes,

which are suggested to use multiple 15-bp direct repeats, the *recer* sequences, within a gene (73,74). This is also different from our case where non-repeated sequences at both the 5'- and 3'-sides were used.

Inactivation of an RM system by insertion of another RM system from a distantly related bacterium

The Type III RM system at locus 4 appeared to have decayed following insertion of a Type II RM system that likely transferred from an oral bacterium (Figure 2). The oral cavity is suggested as a reservoir of *H. pylori* (75), thus frequent interactions between *H. pylori* and oral bacteria may have led to this observation. This replacement of an epigenetic system by another epigenetic system through distant horizontal transfer was likely accompanied by a change in the epigenome, more specifically, in the DNA methylation pattern. This might be a conflict between epigenetic systems (35,36). Other examples of 'RM-on-RM-type' insertions were found in genome comparisons (30,76,77) and are reminiscent of the 'transposon-on-transposon' structure often found in eukaryotic genomes (78,79).

TRD movement between distantly-related bacteria

Homologs of an *H. pylori* TRD homology group were found in a different *Helicobacter* species (Figure 6). Because the *mod* non-TRD sequence belongs to a different homology group, this transfer might have taken place through the above mechanism. In other words, the above mechanism might have promoted distant horizontal transfer of various TRDs.

The *H. pylori* TRD homologs are found in a wide variety of bacteria (Figure 7). We do not yet know the relative contribution of three possible processes to this distribution: TRD movement between *mod* genes, *mod* gene movement and Type III RM system movement.

Biological significance of switches in TRD of RM systems

The capacity to switch the TRD of *mod* gene affects both R and M activity of Type III RM systems. A change in restriction specificity will alter the repertoire of acceptable DNAs. This will not only affect the defense against infecting genetic elements but also, in a wider sense, limit the future direction of genome evolution. The change in the modification specificity will lead to a change in epigenetic methylation states and, therefore, the transcriptome (14,80). The two-sided change could provide the bacterium with an ability to adapt to various environments by changing the genome and global gene expression. Each of these patterns could be modulated by variation in the strength of each enzyme activity. Each of these unique epigenome states may define an elementary unit of natural selection. Such concepts of RM-driven adaptive evolution can be evaluated through examination of additional *H. pylori* genomes and experimentation.

The mobility of *mod* genes in *Helicobacter* species (Figure 6) and the mobility of TRD in distantly related bacteria (Figure 7) are consistent with the concept of epigenetics-driven evolution. They likely lead to changes in the recognition sequence of both R and M genes and therefore, changes in the epigenome. Currently, recognition sequences for Type III *mod* genes of *H. pylori* are known only for HP0260 and HP0593, which are the genes of the 26695 strain with TRD homology group W recognizing GATC at locus 0 and group D recognizing CTGCAG at locus 4 (81). Further experimental determination of recognition sequences by cleavage tests of methylated sites by restriction enzyme with known recognition sequence (81) or by single-molecule real-time sequencing methods (82) would help further understanding of the diversity of RM systems and its biological significance.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–8 and Supplementary Tables 1–3.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Patthy,L. (1999) Genome evolution and the evolution of exon-shuffling—a review. *Gene*, **238**, 103–114.
2. Pingoud,A. (2004) *Restriction Endonucleases*. Springer Verlag, Berlin.
3. Blumenthal,R.M. and Cheng,X. (2002) Restriction-modification systems. In: Yasbin,R. and Streips,U.N. (eds), *Modern Microbial Genetics*. Wiley-Liss, Inc, New York, pp. 177–226.
4. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S., Dryden,D.T.F., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
5. Bujnicki,J.M. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. *BMC Evol. Biol.*, **2**, 3.
6. Malone,T., Blumenthal,R.M. and Cheng,X. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.
7. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell Mol. Life Sci.*, **62**, 685–707.
8. Orlowski,J. and Bujnicki,J.M. (2008) Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. *Nucleic Acids Res.*, **36**, 3552–3569.
9. Murray,N.E. (2000) Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.
10. Gough,J.A. and Murray,N.E. (1983) Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Mol. Biol.*, **166**, 1–19.
11. Dryden,D.T.F., Murray,N.E. and Rao,D.N. (2001) Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.*, **29**, 3728–3741.
12. Fox,K.L., Dowideit,S.J., Erwin,A.L., Srikhanta,Y.N., Smith,A.L. and Jennings,M.P. (2007) *Haemophilus influenzae* phasevarions have evolved from type III DNA restriction systems into epigenetic regulators of gene expression. *Nucleic Acids Res.*, **35**, 5242–5252.
13. Srikhanta,Y.N., Dowideit,S.J., Edwards,J.L., Falsetta,M.L., Wu,H.J., Harrison,O.B., Fox,K.L., Seib,K.L., Maguire,T.L., Wang,A.H. *et al.* (2009) Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. *PLoS Pathog.*, **5**, e1000400.
14. Srikhanta,Y.N., Gorrell,R.J., Steen,J.A., Gawthorne,J.A., Kwok,T., Grimmond,S.M., Robins-Browne,R.M. and Jennings,M.P. (2011) Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. *PLoS One*, **6**, e27569.
15. Betlach,M., Hershfield,V., Chow,L., Brown,W., Goodman,H. and Boyer,H.W. (1976) A restriction endonuclease analysis of the bacterial plasmid controlling the *ecoRI* restriction and modification of DNA. *Fed. Proc.*, **35**, 2037–2043.
16. Dempsey,R.M., Carroll,D., Kong,H., Higgins,L., Keane,C.T. and Coleman,D.C. (2005) *Sau*42I, a *Bcg*I-like restriction-modification

system encoded by the *Staphylococcus aureus*
quadruple-converting phage Phi42. *Microbiology*, **151**, 1301–1311.

17. Hendrix,R.W., Smith,M.C., Burns,R.N., Ford,M.E. and Hatfull,G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.

18. Humbelin,M., Suri,B., Rao,D.N., Hornby,D.P., Eberle,H., Pripfl,T., Kenel,S. and Bickle,T.A. (1988) Type III DNA restriction and modification systems *Eco*P1 and *Eco*P15. Nucleotide sequence of the *Eco*P1 operon, the *Eco*P15 *mod* gene and some *Eco*P1 *mod* mutants. *J. Mol. Biol.*, **200**, 23–29.

19. Kita,K., Kawakami,H. and Tanaka,H. (2003) Evidence for horizontal transfer of the *Eco*T38I restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in *Escherichia coli* TH38 strains. *J. Bacteriol.*, **185**, 2296–2305.

20. Kita,K., Tsuda,J., Kato,T., Okamoto,K., Yanase,H. and Tanaka,M. (1999) Evidence of horizontal transfer of the *Eco*O109I restriction-modification gene to *Escherichia coli* chromosomal DNA. *J. Bacteriol.*, **181**, 6822–6827.

21. Ohshima,H., Matsuoka,S., Asai,K. and Sadaie,Y. (2002) Molecular organization of intrinsic restriction and modification genes *Bsu*M of *Bacillus subtilis* Marburg. *J. Bacteriol.*, **184**, 381–389.

22. Tyndall,C., Lehnherr,H., Sandmeier,U., Kulik,E. and Bickle,T.A. (1997) The type IC *hsd* loci of the enterobacteria are flanked by DNA with high homology to the phage P1 genome: implications for the evolution and spread of DNA restriction systems. *Mol. Microbiol.*, **23**, 729–736.

23. Rochepeau,P., Selinger,L.B. and Hynes,M.F. (1997) Transposon-like structure of a new plasmid-encoded restriction-modification system in *Rhizobium leguminosarum* VF39SM. *Mol. Gen. Genet.*, **256**, 387–396.

24. Siguier,P., Gagnevin,L. and Chandler,M. (2009) The new IS*1595* family, its relation to IS*1* and the frontier between insertion sequences and transposons. *Res. Microbiol.*, **160**, 232–241.

25. Stiens,M., Becker,A., Bekel,T., Godde,V., Goesmann,A., Niehaus,K., Schneiker-Bekel,S., Selbitschka,W., Weidner,S., Schluter,A. *et al.* (2008) Comparative genomic hybridisation and ultrafast pyrosequencing revealed remarkable differences between the *Sinorhizobium meliloti* genomes of the model strain Rm1021 and the field isolate SM11. *J. Biotechnol.*, **136**, 31–37.

26. van Zyl,L.J., Deane,S.M., Louw,L.A. and Rawlings,D.E. (2008) Presence of a family of plasmids (29 to 65 kilobases) with a 26-kilobase common region in different strains of the sulfur-oxidizing bacterium *Acidithiobacillus caldus*. *Appl. Environ. Microbiol.*, **74**, 4300–4308.

27. Bujnicki,J.M. and Radlinska,M. (1999) Molecular phylogenetics of DNA 5mC-methyltransferases. *Acta Microbiol. Pol.*, **48**, 19–30.

28. Jeltsch,A., Kroger,M. and Pingoud,A. (1995) Evidence for an evolutionary relationship among type-II restriction endonucleases. *Gene*, **160**, 7–16.

29. Nobusato,A., Uchiyama,I. and Kobayashi,I. (2000) Diversity of restriction-modification gene homologues in *Helicobacter pylori*. *Gene*, **259**, 89–98.

30. Nobusato,A., Uchiyama,I., Ohashi,S. and Kobayashi,I. (2000) Insertion with long target duplication: a mechanism for gene mobility suggested from comparison of two related bacterial genomes. *Gene*, **259**, 99–108.

31. 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) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.

32. Furuta,Y., Kawai,M., Yahara,K., Takahashi,N., Handa,N., Tsuru,T., Oshima,K., Yoshida,M., Azuma,T., Hattori,M. *et al.* (2011) Birth and death of genes linked to chromosomal inversion. *Proc. Natl Acad. Sci. USA*, **108**, 1501–1506.

33. Kusano,K., Sakagami,K., Yokochi,T., Naito,T., Tokinaga,Y., Ueda,E. and Kobayashi,I. (1997) A new type of illegitimate recombination is dependent on restriction and homologous interaction. *J. Bacteriol.*, **179**, 5380–5390.

34. Handa,N., Nakayama,Y., Sadykov,M. and Kobayashi,I. (2001) Experimental genome evolution: large-scale genome

rearrangements associated with resistance to replacement of a chromosomal restriction-modification gene complex. *Mol. Microbiol.*, **40**, 932–940.

35. Ishikawa,K., Fukuda,E. and Kobayashi,I. (2010) Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. *DNA Res.*, **17**, 325–342.

36. Fukuda,E., Kaminska,K.H., Bujnicki,J.M. and Kobayashi,I. (2008) Cell death upon epigenetic genome methylation: a novel function of methyl-specific deoxyribonucleases. *Genome Biol.*, **9**, R163.

37. Ishikawa,K., Handa,N. and Kobayashi,I. (2009) Cleavage of a model DNA replication fork by a Type I restriction endonuclease. *Nucleic Acids Res.*, **37**, 3531–3544.

38. Asakura,Y. and Kobayashi,I. (2009) From damaged genome to cell surface: transcriptome changes during bacterial cell death triggered by loss of a restriction-modification gene complex. *Nucleic Acids Res.*, **37**, 3021–3031.

39. Mochizuki,A., Yahara,K., Kobayashi,I. and Iwasa,Y. (2006) Genetic addiction: selfish gene's strategy for symbiosis in the genome. *Genetics*, **172**, 1309–1323.

40. Suerbaum,S. and Michetti,P. (2002) *Helicobacter pylori* infection. *N. Engl. J. Med.*, **347**, 1175–1186.

41. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.

42. Xu,Q., Morgan,R.D., Roberts,R.J. and Blaser,M.J. (2000) Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc. Natl Acad. Sci. USA*, **97**, 9671–9676.

43. Falush,D., Kraft,C., Taylor,N.S., Correa,P., Fox,J.G., Achtman,M. and Suerbaum,S. (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl Acad. Sci. USA*, **98**, 15056–15061.

44. Dubnau,D. (1999) DNA uptake in bacteria. *Annu. Rev. Microbiol.*, **53**, 217–244.

45. Linz,B., Balloux,F., Moodley,Y., Manica,A., Liu,H., Roumagnac,P., Falush,D., Stamer,C., Prugnolle,F., van der Merwe,S.W. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, **445**, 915–918.

46. Furuta,Y., Kawai,M., Uchiyama,I. and Kobayashi,I. (2011) Domain movement within a gene: A novel evolutionary mechanism for protein diversification. *PLoS One*, **6**, e18819.

47. Kawai,M., Furuta,Y., Yahara,K., Tsuru,T., Oshima,K., Handa,N., Takahashi,N., Yoshida,M., Azuma,T., Hattori,M. *et al.* (2011) Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol.*, **11**, 104.

48. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

49. Reiser,J. and Yuan,R. (1977) Purification and properties of the P15 specific restriction endonuclease from Escherichia coli. *J. Biol. Chem.*, **252**, 451–456.

50. Meisel,A., Kruger,D.H. and Bickle,T.A. (1991) M.EcoP15 methylates the second adenine in its recognition sequence. *Nucleic Acids Res.*, **19**, 3997.

51. Davies,G.P., Martin,I., Sturrock,S.S., Cronshaw,A., Murray,N.E. and Dryden,D.T.F. (1999) On the structure and operation of type I DNA restriction enzymes. *J. Mol. Biol.*, **290**, 565–579.

52. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

53. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

54. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.

55. Yarza,P., Richter,M., Peplies,J., Euzeby,J., Amann,R., Schleifer,K.H., Ludwig,W., Glockner,F.O. and Rossello-Mora,R. (2008) The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.*, **31**, 241–250.

56. Tamura,K., Peterson,D., Peterson,N., Stecher,G., Nei,M. and Kumar,S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.

57. Nakamura,Y., Itoh,T., Matsuda,H. and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.*, **36**, 760–766.

58. de Vries,N., Duinsbergen,D., Kuipers,E.J., Pot,R.G., Wiesenekker,P., Penn,C.W., van Vliet,A.H., Vandenbroucke-Grauls,C.M. and Kusters,J.G. (2002) Transcriptional phase variation of a type III restriction-modification system in Helicobacter pylori. *J. Bacteriol.*, **184**, 6615–6623.

59. Moore,W.E., Holdeman,L.V., Smibert,R.M., Hash,D.E., Burmeister,J.A. and Ranney,R.R. (1982) Bacteriology of severe periodontitis in young adult humans. *Infect. Immun.*, **38**, 1137–1148.

60. Fischer,W., Windhager,L., Rohrer,S., Zeiller,M., Karnholz,A., Hoffmann,R., Zimmer,R. and Haas,R. (2010) Strain-specific genes of Helicobacter pylori: genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Res.*, **38**, 6089–6101.

61. Srikhanta,Y.N., Fox,K.L. and Jennings,M.P. (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.*, **8**, 196–206.

62. Sharp,P.M., Kelleher,J.E., Daniel,A.S., Cowan,G.M. and Murray,N.E. (1992) Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. *Proc. Natl Acad. Sci. USA*, **89**, 9836–9840.

63. Fuller-Pace,F.V., Bullas,L.R., Delius,H. and Murray,N.E. (1984) Genetic recombination can generate altered restriction specificity. *Proc. Natl Acad. Sci. USA*, **81**, 6095–6099.

64. Gann,A.A., Campbell,A.J., Collins,J.F., Coulson,A.F. and Murray,N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Mol. Microbiol.*, **1**, 13–22.

65. Gubler,M., Braguglia,D., Meyer,J., Piekarowicz,A. and Bickle,T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. *EMBO J.*, **11**, 233–240.

66. Chen,J.M., Cooper,D.N., Chuzhanova,N., Ferec,C. and Patrinos,G.P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.*, **8**, 762–775.

67. Alm,R.A., Bina,J., Andrews,B.M., Doig,P., Hancock,R.E. and Trust,T.J. (2000) Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infect. Immun.*, **68**, 4155–4168.

68. Yamaoka,Y. and Alm,R.A. (2008) Helicobacter pylori outer membrane proteins. In: Yamaoka,Y. (ed.), *Helicobacter Pylori: Molecular Genetics and Cellular Biology*. Caister Academic Pr, Norfolk, pp. 37–60.

69. Haas,R. and Meyer,T.F. (1986) The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. *Cell*, **44**, 107–115.

70. Brayton,K.A., Palmer,G.H., Lundgren,A., Yi,J. and Barbet,A.F. (2002) Antigenic variation of Anaplasma marginale msp2 occurs by combinatorial gene conversion. *Mol. Microbiol.*, **43**, 1151–1159.

71. Barbour,A.G. (1990) Antigenic variation of a relapsing fever Borrelia species. *Annu. Rev. Microbiol.*, **44**, 155–171.

72. Tsuru,T. and Kobayashi,I. (2008) Multiple genome comparison within a bacterial species reveals a unit of evolution spanning two adjacent genes in a tandem paralog cluster. *Mol. Biol. Evol.*, **25**, 2457–2473.

73. de Chateau,M. and Bjorck,L. (1996) Identification of interdomain sequences promoting the intronless evolution of a bacterial protein family. *Proc. Natl Acad. Sci. USA*, **93**, 8490–8495.

74. Rich,R.L., Kreikemeyer,B., Owens,R.T., LaBrenz,S., Narayana,S.V., Weinstock,G.M., Murray,B.E. and Hook,M. (1999) Ace is a collagen-binding MSCRAMM from Enterococcus faecalis. *J. Biol. Chem.*, **274**, 26939–26945.

75. Zaric,S., Bojic,B., Jankovic,L., Dapcevic,B., Popovic,B., Cakic,S. and Milasin,J. (2009) Periodontal therapy improves gastric Helicobacter pylori eradication. *J. Dent. Res.*, **88**, 946–950.

76. Lin,L.F., Posfai,J., Roberts,R.J. and Kong,H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA*, **98**, 2740–2745.

77. Furuta,Y., Abe,K. and Kobayashi,I. (2010) Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res.*, **38**, 2428–2443.

78. Pace,J.K. 2nd and Feschotte,C. (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.*, **17**, 422–432.

79. Jiang,N. and Wessler,S.R. (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell*, **13**, 2553–2564.

80. Srikhanta,Y.N., Maguire,T.L., Stacey,K.J., Grimmond,S.M. and Jennings,M.P. (2005) The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. *Proc. Natl Acad. Sci. USA*, **102**, 5547–5551.

81. Vitkute,J., Stankevicius,K., Tamulaitiene,G., Maneliene,Z., Timinskas,A., Berg,D.E. and Janulaitis,A. (2001) Specificities of eleven different DNA methyltransferases of *Helicobacter pylori* strain 26695. *J. Bacteriol.*, **183**, 443–450.

82. Clark,T.A., Murray,I.A., Morgan,R.D., Kislyuk,A.O., Spittle,K.E., Boitano,M., Fomenkov,A., Roberts,R.J. and Korlach,J. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.*, **40**, e29.