4046–4059 *Nucleic Acids Research, 2006, Vol. 34, No. 14* *Published online 16 August 2006*
doi:10.1093/nar/gkl568

High allelic diversity in the methyltransferase gene of a phase variable type III restriction-modification system has implications for the fitness of *Haemophilus influenzae*

Christopher D. Bayliss*, Martin J. Callaghan[1] and E. Richard Moxon

Molecular Infectious Diseases Group, Weatherall Institute for Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK and [1]Department of Paediatrics, University of Oxford, John Radcliffe Hospital, Level 4, Headington, Oxford, OX3 9DU, UK

Received June 13, 2006; Revised July 21, 2006; Accepted July 22, 2006

ABSTRACT

Phase variable restriction-modification (R-M) systems are widespread in Eubacteria. *Haemophilus influenzae* encodes a phase variable homolog of Type III R-M systems. Sequence analysis of this system in 22 non-typeable *H.influenzae* isolates revealed a hypervariable region in the central portion of the *mod* gene whereas the *res* gene was conserved. Maximum likelihood (ML) analysis indicated that most sites outside this hypervariable region experienced strong negative selection but evidence of positive selection for a few sites in adjacent regions. A phylogenetic analysis of 61 Type III *mod* genes revealed clustering of these *H.influenzae mod* alleles with *mod* genes from pathogenic *Neisseriae* and, based on sequence analysis, horizontal transfer of the *mod–res* complex between these species. Neisserial *mod* alleles also contained a hypervariable region and all *mod* alleles exhibited variability in the repeat tract. We propose that this hypervariable region encodes the target recognition domain (TRD) of the Mod protein and that variability results in alterations to the recognition sequence of this R-M system. We argue that the high allelic diversity and phase variable nature of this R-M system have arisen due to selective pressures exerted by diversity in bacteriophage populations but also have implications for other fitness attributes of these bacterial species.

INTRODUCTION

Restriction-modification (R-M) systems, ubiquitous in eubacteria and archaea, are thought to protect organisms from invasion by foreign DNA. Phase variation (PV), or reversible, high frequency changes in gene expression, is a feature of many R-M systems and phase variable R-M systems are found in a variety of species including *Helicobacter pylori*, *Mycoplasma pulmonis*, *Pasteurella hemolytica*, *Neisseria meningitidis* and *Haemophilus influenzae* (1–5). Surprisingly, information on these phase variable R-M systems is limited even though phase variable expression of DNA modification or restriction is likely to have important implications for the evolutionary fitness of many bacterial species.

Most R-M systems comprise a DNA methyltransferase (MTase) and a restriction endonuclease. The MTase enables recognition of 'self' DNA by methylation of specific nucleotides within particular DNA sequences whereas the endonuclease enzymatically cleaves 'foreign' unmodified DNA (e.g. the genome of an invading bacteriophage). R-M systems are classified into three major groups—Types I, II and III—and one minor group—Type IV (6). Type I systems are encoded by three genes, *hsdM*, *hsdS* and *hsdR*, whose products form two multi-subunit enzymes; a DNA MTase composed of HsdM and HsdS sub-units; and an endonuclease composed of HsdM, HsdS and HsdR sub-units (7). Type II systems, of which there are more than 3500 (6), consist of two genes, M and R, whose products are an MTase and an endonuclease (8). Type III systems are comprised of two genes, *mod* and *res*, whose products form two multi-subunit enzymes: a MTase composed only of Mod sub-units; and an endonuclease composed of both Mod and Res sub-units (9).

*To whom correspondence should be addressed at: Molecular Bacteriology and Immunology Group, Institute for Infection, Immunity and Inflammation, School of Molecular Medical Sciences, University of Nottingham, Queens Medical Centre, NG7 2UH, Nottingham. Tel: +44 115 8230743; Fax: +44 115 8230759; Email: mrzcb1@gwmail.nottingham.ac.uk

MTases of all four Types of R-M systems can be classified according to the organization of 10 conserved amino acid motifs (10,11). These motifs form a sub-domain of the active site and the binding domain for the methylation substrate *S*-adenosyl-L-methionine. Recent observations of a common structural topology known as an 'MTase fold' suggests that the MTase's of R-M systems may have evolved from other Rosmann-fold proteins (11). The MTases of Type I and III R-M systems have a gamma or beta-organization of the conserved motifs and form distinct clusters in phylogenetic analyses incorporating MTases of Types I, II and III (12). R-M systems exhibit extensive diversity in their DNA sequence recognition specificities. In Type II R-M systems, this specificity is conferred by an autonomous target recognition domain (TRD), which is present in both the MTase and the endonuclease enzymes enabling independent binding of each enzyme to the recognition site (11). In Type I and III R-M systems binding specificity is conferred by the HsdS and Mod sub-units, respectively. Strikingly, *hsdS* genes exhibit high levels of divergence within individual species and can undergo recombinatorial re-assortment of the two DNA-binding domains resulting in enzymes with altered recognition sites (7,13). The TRDs of Type II systems also exhibit significant diversity but TRDs for Type III MTases have not been characterized.

Various hypotheses have been advanced for the evolution of R-M systems in general and phase variable R-M systems in particular. The ubiquity of Type II R-M systems is often ascribed to either a role in defence against bacteriophage infection or to their exhibition of selfish characteristics (14,15). In support of the latter theory, inactivation of the MTase subunit is lethal due to restriction of the genome by the endonuclease. In contrast, inactivation of the MTase genes of Type I systems is not lethal due in part to preferential degradation of the endonuclease by the protease ClpXP, such that restriction activity is lost prior to methylation activity (7,15). Although, the specific mechanisms for control of Type III restriction endonucleases are unknown, the observation that many Type III R-M systems, as well as numerous Type I but not Type II R-Ms, are subject to PV (16) indicates that loss of expression of a Type III R-M is not lethal. As noted by Dybvig and co-workers for a phase variable Type I R-M systems (13), an ability to phase vary is not consistent with the concept of 'selfish elements'. Another explanation for the evolution of phase variable R-M systems is that, there is fitness benefits associated with the temporary alleviation of the barrier imposed by these systems to the acquisition of foreign, and potentially beneficial, DNA molecules in naturally transformable species, such as *H.pylori* (17–19). Alternatively it has been proposed that these systems influence the expression or PV rates of other genes permitting adaptation to changes in environmental conditions by altering expression of unlinked genes (20,21).

H.influenzae is an obligate commensal of the upper respiratory tract of humans whose genome contains multiple phase variable genes (22). Two of the phase variable genes, also termed contingency loci (23), of *H.influenzae* strain Rd encode sub-units of known or putative R-M systems. One of these loci encodes HindI, a Type I R-M system (24,25). The *hsdM* gene of this system contains a pentanucleotide repeat tract (5′-GAGAC) whose numbers vary between strains and during persistence of strains in cystic fibrosis patients (26). Alterations in repeat number mediate phase variable resistance to infection with HP1c1, a *H.influenzae* bacteriophage (24). The second locus contains a *mod* (HI1056) and a *res* (HI1055) gene with homology to Type III R-M systems (1). Type III R-M enzymes, HinfI and HineI, were purified from *H.influenzae* strains Rf and Re, respectively, and shown to methylate adenine in the sequence 5′-GCAAT (27–29). Although the genes encoding these systems were not identified, there is only one Type III R-M system present in the two published *H.influenzae* genome sequences (see www.tigr.org) suggesting that HinfI and HineI are encoded by alleles of HI1056 and HI1055. A 5′-AGTC repeat tract present in *mod* mediates high frequency PV from this locus (1). Recently, PV of the *mod* gene of *H.influenzae* strain Rd was found to alter the expression of a number of unlinked genes, including an outer membrane protein and heat shock proteins, suggesting that this locus constituted a 'phasevarion', i.e. a stochastically expressed transcriptional regulon, whose expression generates fitness benefits in specific environments (20).

Despite a wide distribution, few phase variable R-M systems have been studied in detail. The aim of this study was to investigate the molecular evolution of a phase variable Type III R-M system of *H.influenzae*, which has been implicated as having a major influence on the fitness of this species. Sequence analysis of the *mod* and *res* genes of this R-M system from a number of isolates was conducted and a hypervariable region was observed in the central portion of *mod*. A likelihood based analysis of evolutionary selection pressures revealed the effect of diversifying selection pressure on this gene. Additionally, statistical tests of recombination, phylogenetic analyses and visual comparisons were employed to determine the extent of horizontal genetic exchange of *mod* sequences. A laterally-transferred homolog of this system in the pathogenic *Neisseriae* was also observed suggesting the existence of a common gene pool. Finally we speculate on the implications of our findings for the evolution of novel R-M specificities and their exchange among bacterial species with differing pathogenicities and for the proposed biological roles of phase variable R-M systems.

MATERIALS AND METHODS

Strains

Non-typeable *H.influenzae* (NT*Hi*) isolates were supplied by J. Eskola and the Finnish Otitis Media Study Group. These isolates were collected from children with otitis media. Isolates were grown in brain heart infusion (BHI) supplemented with either haemin (10 μg/ml) or NAD (2 μg/ml) for liquid media or Levinthal supplement (10%) for solid media.

Amplification and sequencing of R-M locus

Oligonucleotide primers for PCR amplification and sequencing of the R-M genes were initially designed using the strain Rd genome sequence. Subsequently, isolate specific primers were designed using the DNA sequences derived for an isolate. Oligonucleotide primer sequences are provided in Supplementary Table 1.

Multiple over-lapping PCR fragments were generated for each isolate and these fragments were used in DNA sequencing reactions. The complete *mod* and *res* gene sequences and the partial *mod* gene sequences presented in Supplementary Figure 1 were generated from two or more DNA sequence reactions covering both DNA strands. The partial *mod* sequences of isolates 981, 477, 1268, 1200, 1231, 1232, 1158, 1159, 486, 1209, 375, 1247, 1180, 1181 and 1292 that extended into the hypervariable region (and presented in Supplementary Figure 2) were derived from only one DNA strand.

DNA sequencing reactions were performed using the ABI PRISM BigDye terminator cycle sequencing kit (Perkin-Elmer) and electrophoresed on 4.25% gels using an ABI 377 Automated Sequencer. The sequence traces specific for each isolate were aligned and edited using the PreGap and Gap4 programs (Staden).

Genomic sequences for *mod* and *res* from *H.influenzae* strains Rd, R2846 and R2866 were obtained through the TIGR Microbial database (www.tigr.org) and the NCBI databases (www.ncbi.nlm.nih.gov). These sequences were aligned with those obtained herein using Pileup (Wisconsin Package Version 10.0, Genetics Computer Group, Madison, WI). GeneDoc version 2.6.001 (30) (www.psc.ed/biomed/genedoc) was employed for further editing and generation of figures. Where appropriate the numbers of repeats were altered to obtain full-length amino acid sequences.

Selection tests

A nucleotide sequence alignment of the semi-conserved regions of *mod* genes from the 22 NT*Hi* isolates and strain Rd was constructed. The alignment was based on the translated amino acid sequences of these regions and was constructed manually. A maximum likelihood (ML) phylogeny of the nucleotide sequences was then generated in the program PAUP* (31) using the GTR model of nucleotide substitution, with the value for transition/transversion rate and the shape parameter of an eight-category gamma distribution of rate variation among sites, estimated from the data. The evolutionary selection pressure at each codon site was investigated by ML analysis as described previously (32). The program CODEML, part of the PAML package version 3.0 (33), was employed to compare the likelihoods of a series of six models of codon substitution (each of which assumed different types of selection). The log likelihood values for sets of models, as appropriate, were compared using the likelihood ratio test. The significance of the results was tested by Chi-squared analysis, with the number of degrees of freedom equal to the difference in the number of parameters between models. Positive selection was inferred when the log likelihood of the models allowing for positive selection (models M2, M3 and M8) were significantly favoured in comparison to those which did not (Models M0, M1 and M7), and when the d_N/d_S ratio at individual codon sites in M2, M3 and M8 was significantly greater than 1. Full ML parameters are available from the authors upon request.

Phylogenetic analysis

Nucleotide and amino acid sequences for other Type III R-M MTase sequences were obtained from the TIGR Microbial database, the NCBI databases or from REBASE (rebase.neb.com). Identification of Type III R-M systems relied on annotations provided in the TIGR database and REBASE. Sequences with out-of-frame numbers of repeats were manually-altered to obtain in-frame numbers of repeats and full-length amino acid sequences. Amino acid sequences were aligned and a number of semi-conserved regions were identified visually. These regions were then aligned separately and subjected to phylogenetic analysis using the neighbour joining (NJ) algorithm implemented in the program MEGA, version 3 (34). This algorithm clustered sequences together by amino acid *p*-distance and used a pairwise deletion method to compensate for alignment gaps among sequences. Alignments of a sub-group of the MTases (see text) were generated using amino acid sequences lacking the repeat tracts and variable regions. Phylogenetic trees of these alignments were constructed as for the semi-conserved regions.

Recombination analysis

Recombination was investigated in amino acid and nucleotide sequence alignments using default parameters in the program GENECONV version 1.81 [S.A.Sawyer (1999) GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at www.math.wustl.edu/~sawyer/]. Recombination was inferred when the simulated *P*-values for global inner recombination fragments were found to be significant ~95%. The significance of global inner fragments was also tested by Bonferroni corrected Karlin–Altschul *P*-values, a more conservative, but approximate, test of significance.

Nomenclature of *mod* types

The different alleles of *mod* are referred to in this paper with reference to the isolate from which the sequences were derived (e.g. 375*mod*). The gene products were given a reference related to the sequence of the variable region (i.e. Mod types) with each name being followed by a number relating to the Mod type (e.g. 375*mod* encodes M.Hin1056ModP-2; see Table 1). In each case, except one, the Mod types were identical at the sequence level. The exception was Mod type 7 for which sequences were found in isolates 162 and 486. These sequences were identical within the variable region (hence Mod type 7) but differed outside this region and so were designated A and B. The 1056 in each name refers to gene HI1056 (Figure 1) in the strain Rd genome sequence, which encodes the prototype gene for this R-M system. The REBASE designations are based on including the name of the strain in which the gene is found and, if from a genome sequence, the name or number of the open reading frame (ORF).

Accession numbers

The GenBank accession numbers for the novel sequences included in this paper are: 432*mod*, DQ857350; 1233*mod*, DQ857351; 285*mod*, DQ857352; 1008*mod*, DQ857353; 1124*mod*, DQ857354; 162*mod*, DQ857355; 667*mod*, DQ857356; 375*mod*, DQ857357; 486*mod*, DQ857358; 981*mod*, DQ857359; 1158*mod*, DQ857360; 1159*mod*, DQ857361; 1180*mod*, DQ857362; 1181*mod*, DQ857363; 1200*mod*, DQ857364, 1209*mod*, DQ857365; 1231*mod*,

DQ857366; 1232*mod*, DQ857367; 1247*mod*, DQ857368; 1292*mod*, DQ857369; 1268*mod*, DQ857370; 477*mod*, DQ857371.

RESULTS

Identification of sequence diversity in multiple *H.influenzae* mod genes

Genes HI1056 and HI1055 of *H.influenzae* strain Rd have homology to MTase and endonuclease enzymes, respectively, of Type III R-M systems (1). These genes are termed *hindorf1056MP* and *hindorf1056RP*, respectively, in REBASE [rebase.neb.com, (35)] but will be referred to as *mod* and *res* herein (similarly, the gene products, M.Hindorf1056P and Hindorf1056P, will be referred to as Mod and Res). The *mod* gene of strain Rd contains a tract of 32 5′-AGTC repeats. Previously, this repeat tract was found to exhibit significant diversity in a range of NT*Hi* isolates [Table 1, (1)]. In order to investigate the degree of conservation of other regions of *mod*, we performed an extensive sequence analysis of the upstream sequences and N-terminal region of this gene in 22 of these NT*Hi* isolates. A sub-set of isolates, representative of the diversity detected in this initial survey, were then selected for generation of full-length *mod* sequences and an in depth molecular genetic analysis.

Table 1. Repeat numbers and *mod* types for *H.influenzae* and *N.meningitidis* strains and isolates

Strain/isolate designation[a]	MLST type[b]	Repeat number[c]	Initiation codon[d]	Mod type	*mod* N-terminus % similarity to Rd[e]	*mod* full gene % similarity to Rd[e]	*res* full gene % similarity to Rd[e]	Simple designation (REBASE designations[f])
Rd	47	32	Proximal	1	100	100	100	M.Hin1056ModP-1 (M.HindORF1056P)
86-028	—	16	Distal	2	94	79	94	M.Hin1056ModP-2 (M.Hin86ORF1217P,
375	3	16	Distal	2	94	ND[g]	ND	M.Hin375ORFAP,
432	40	16	Distal	2	94	79	ND	M.Hin432ORFAP,
1124	12	16	Distal	2	94	79	94	M.Hin1124ORFAP,
1247	33	19	Distal	2	94	ND	ND	M.Hin1247ORFAP)
667	57	0	Distal	3	94	80	92	M.Hin1056ModP-3 (M.Hin667ORFAP,
1180	2	0	Distal	3	94	ND	ND	M.Hin1180ORFAP,
1181	2	0	Distal	3	94	ND	ND	M.Hin1181ORFAP,
1292	2	0	Distal	3	94	ND	ND	M.Hin1292ORFAP)
1231	34	2	Proximal	4	95	ND	ND	M.Hin1056ModP-4 (M.Hin1231ORFAP,
1232	34	2	Proximal	4	95	ND	ND	M.Hin1232ORFAP,
R2846	257	3	None	4	95	78	95	M.Hin2846ORFAP)
285	39	7	None	5	94	78	ND	M.Hin1056ModP-5 (M.Hin285ORFAP,
477	1	14	Distal	5	94	ND	ND	M.Hin477ORFAP,
981	42	8	Distal	5	94	ND	ND	M.Hin981ORFAP,
1200	36	10	None	5	94	ND	ND	M.Hin1200ORFAP,
1268	36	17	Distal	5	94	ND	ND	M.Hin1268ORFAP)
1158	11	0	Distal	6	93	ND	ND	M.Hin1056ModP-6 (M.Hin1158ORFAP
1159	11	0	Distal	6	93	ND	ND	M.Hin1159ORFAP)
486	41	0	Distal	7A	94	ND	ND	M.Hin1056ModP-7A (M.Hin486ORFAP)
162	37	0	Distal	7B	95	81	96	M.Hin1056ModP-7B (M.Hin162ORFAP)
1008	43	0	Distal	8	94	77	ND	M.Hin1056ModP-8 (M.Hin1008ORFAP)
1209	13	19	Distal	9	92	ND	ND	M.Hin1056ModP-9 (M.Hin1209ORFAP,
1233	13	5	Proximal	9	92	77	ND	M.Hin1233ORFAP)
R2866	99	16	Distal	10	95	83	94	M.Hin1056ModP-10 (M.Hin2866ORFAP)
MC58	—	20	Proximal	11	93	79	91	M.Nme1056ModP-11 (M.NmeBORF1375P)
Z2491	—	3	None	12	95	77	92	M.Nme1056ModP-12 (M.NmeAORF1590P)
FA1090	—	37	Distal	13	92	76	ND	M.Ngo1056ModP-13 (M.NgoORFC707P)

[a]NT*Hi* isolate numbers in normal type. Designations of strains for which genomic data were obtained are in italics.
[b]From *H.influenzae* MLST database at http://haemophilus.mlst.net.
[c]All 5′-AGCC except strain Rd which has 5′-AGTC.
[d]Distal initiation codon is the predicted initiation codon (5′-ATG) of HI1058. Proximal initiation codon is an 5′-ATG that is 53 bp upstream of the repeat tract.
[e]Comparison of nucleotide sequences of isolate to sequence from strain Rd. N-terminal region includes sequences from the distal initiation codon to the repeat tract, the repeat tract and the 264 bp downstream of the repeat tract.
[f]REBASE designations were developed in consultation with Richard Roberts.
ND = no data.

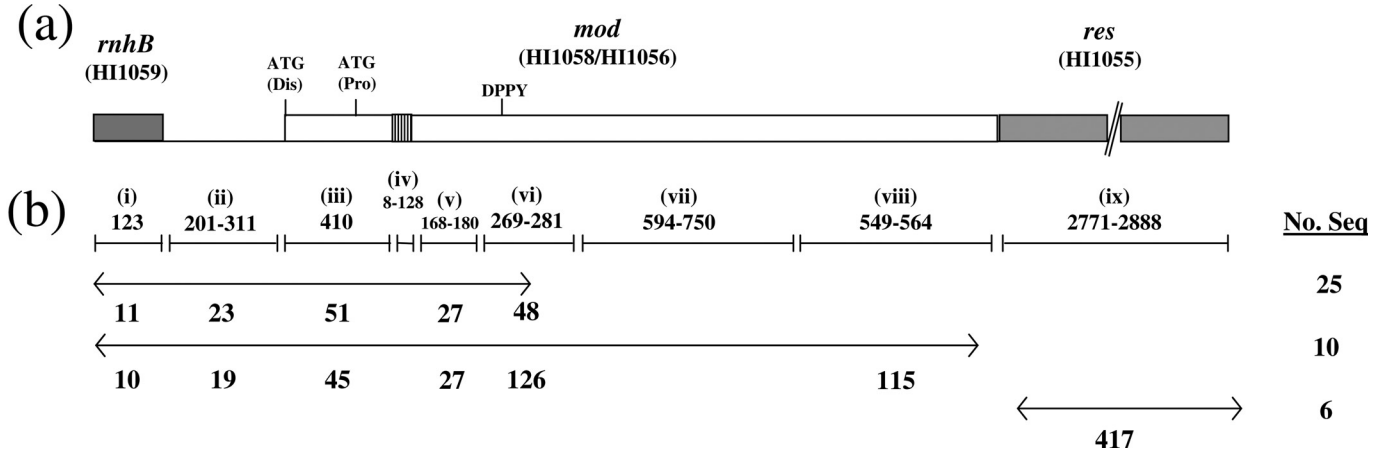
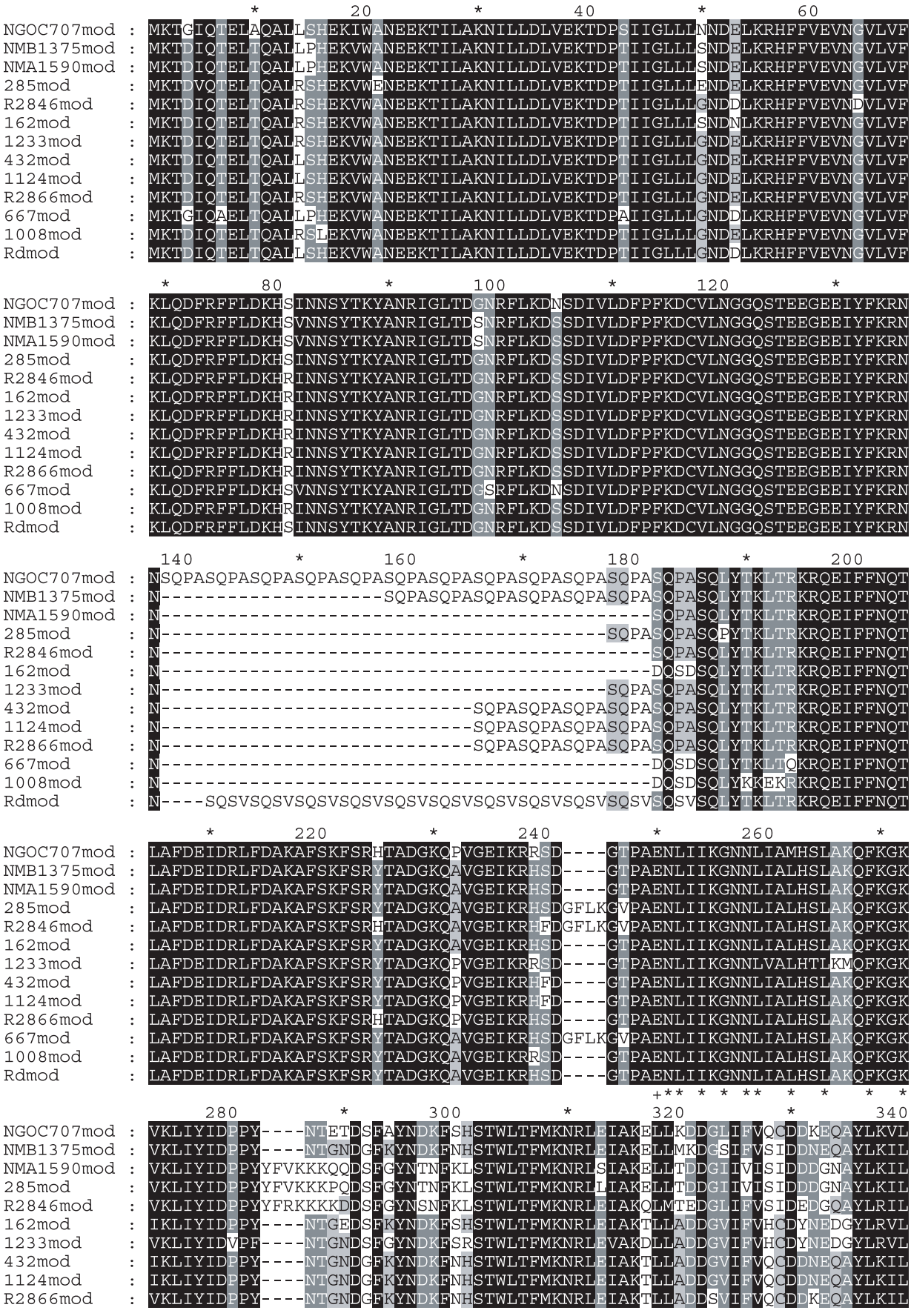


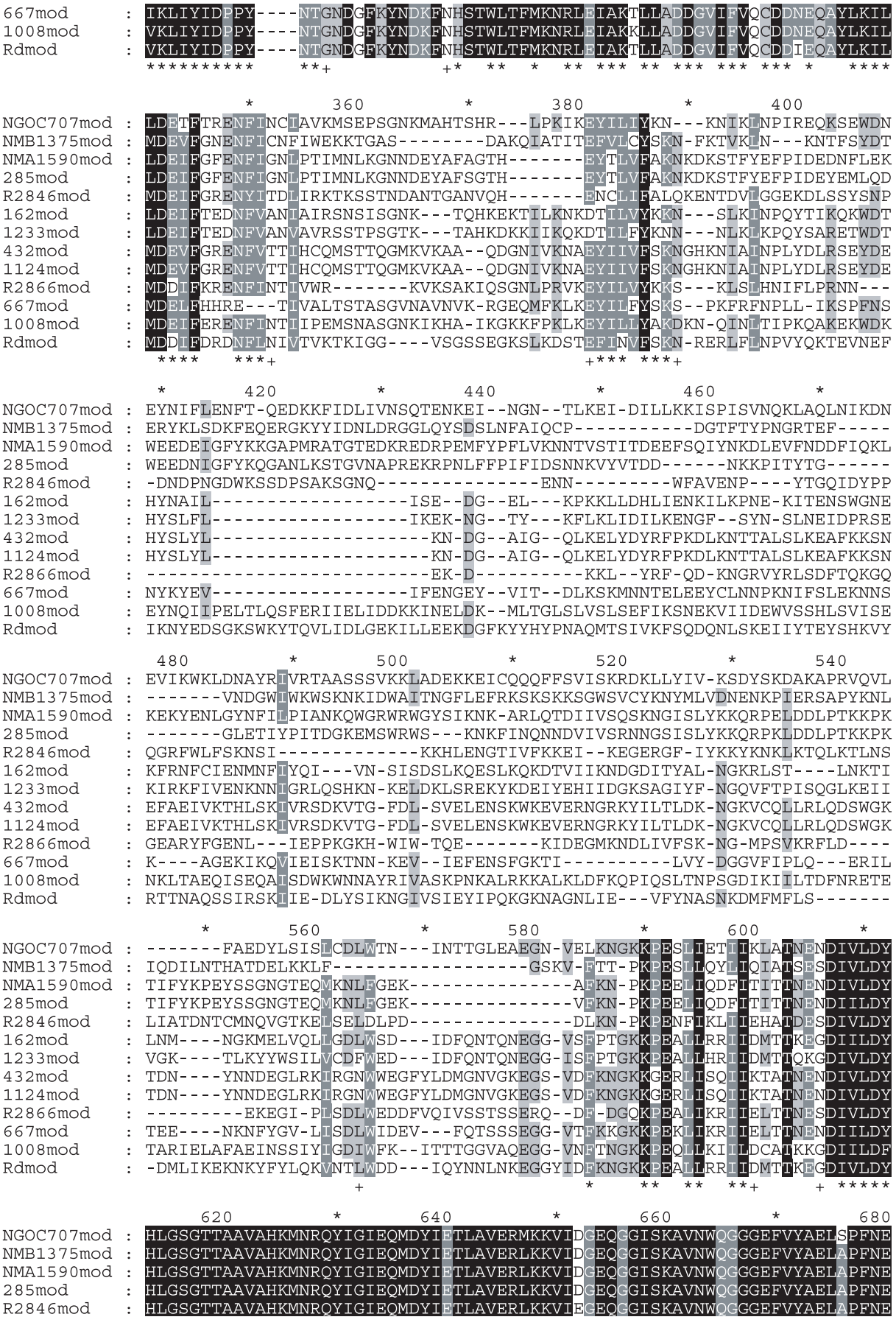
Figure 1. Schematic diagram of the phase variable Type III R-M system of *H.influenzae* and a summary of the allelic variation of *H.influenzae* isolates in this locus. Panel (**a**) shows the reading frames as open or filled rectangles. The intergenic region is shown as a line. The striped box in *mod* marks the position of the repeat tract whilst the diagonal lines in *res* signify that the full gene is not represented in the figure. The HI numbers are from the annotation of the *H.influenzae* strain Rd genome sequence (www.tigr.org). The relative positions of the distal (Dis) and proximal (Pro) 5′-ATG initiation codons and of the DPPY motif, characteristic of many R-M systems, are also marked. The top diagram of panel (**b**) specifies the variations in length (due to insertions/deletions), in nucleotides, of different regions of this locus. These regions are: (i) *rnhB*; (ii) intergenic sequences; (iii) sequences between distal initiation codon and repeat tract; (iv) repeat tract; (v) conserved sequences; (vi) region containing N-terminal MTase motifs; (vii) variable region; (viii) region containing C-terminal MTase motifs; (ix) *res*. The lower diagram indicates the extent of the partial sequences of the *mod* locus (upper line) and of the full-length sequences of *mod* (middle line) and *res* (lower line). The numbers below these lines signify the polymorphic sites in each region observed in comparisons of multiple sequences with the number of *H.influenzae* isolates (see text) analysed being listed on the right-hand side of the figure.

The twenty-two NT*Hi* isolates were chosen from across a ribotyping dendogram containing >400 *H.influenzae* isolates and are representative of the diversity of this species (36). These 22 isolates were obtained from patients with otitis media and included isolates from different ears of the same patient (1158 and 1159, 1231 and 1232, 1180 and 1181), which provided controls for the accuracy of the sequence analysis and evidence for alterations in repeat number (i.e. PV) during the course of infection. The region around the repeat tract of *mod* was amplified from chromosomal DNA using specific primers and the products subjected to direct sequencing using both PCR primers and internal primers. These nucleotide sequences were then compared with those from three genome sequences (strains Rd, R2866 and R2864; Supplementary Figure 1). The sequences started within gene HI1059 (*rnhB*) which was conserved with 11 polymorphic sites in 123 nt (9%; region i in Figure 1). The intergenic region between this gene and *mod* also exhibited few polymorphic sites (23 sites, 7%; region ii in Figure 1) but contained four small insertions (10, 1, 2 and 3 bp in 7, 5, 2 and 3 strains, respectively) and one large 109 bp insertion (4 strains), emphasising a plasticity that may be expected of non-coding regions. Potential promoter elements (−35 5′-TAAATATA; −10 5′-TGTAAT) and a strong Shine–Delgarno sequence (5′-AGGA) were conserved close to the distal initiation codon in all strains. The ORF between the distal initiation codon of *mod* and the repeat tract (when present) was conserved with 51 polymorphic sites in 410 nt (12%; region iii in Figure 1). Similarly, the first 180 nt 3′ to the repeat tract were also conserved with 27 polymorphic sites (15%; region v in Figure 1) and a 12 nt insertion (Supplementary Figure 1). Ten of the polymorphic sites were non-synonymous and altered the amino acid sequence. The next 126 nt encoded amino acid sequences with significant homology to other Type III MTases and included the motif DPPY (see Figure 2) characteristic of many R-M systems. This region had 48 polymorphic sites (part of region vi in Figure 1) with 16 of these being in the final 27 nt (Supplementary Figure 1). Thirteen of the polymorphic sites produced non-synonymous changes. In two isolates, 1209 and 1233, the DPPY motif was altered to DVPF (amino acid positions 236 and 238, Supplementary Figure 2). The polymorphic sites at the end of these sequences were used to define Mod types (Table 1). All sequences of each Mod type were identical apart from those of Mod type 3. In this case a polymorphism 14 nt downstream of the distal initiation codon separates these isolates into two groups (477 and 981 versus 285, 1200 and 1268) whilst a further polymorphism in the *rnhB* gene separates isolates 477 and 981.

Nucleotide sequence data was then obtained from one DNA strand that extended up to 450 nt further into the central portion of the gene. Comparisons indicated significant levels of sequence variability to the extent that the different sequences could not be aligned (Supplementary Figure 3). A similar level of variability was exhibited by the amino acid sequences (data not shown). Notably the nucleotide sequences of the variable region were highly conserved for different isolates of the same Mod type. The alignments of these sequences indicated that the central portion of *mod* was hypervariable.

In order to establish the extent of this variability, complete sequences were obtained for *mod* from seven strains and compared to those derived from genome sequences. Two alleles (from isolates 432 and 1124) of Mod type 2 were sequenced and were found to be identical throughout their length with each other and with that of the *mod* gene from the genome sequence of strain 86-028 (www.tigr.org), i.e. M.Hin86orf1217P in REBASE. Two of the genome sequences (Rd and R2866) contained unique Mod types whilst the *mod* sequence of strain R2846 was identical,


NGOC707mod
NMB1375mod
NMA1590mod
285mod
R2846mod
162mod
1233mod
432mod
1124mod
R2866mod
667mod
1008mod
Rdmod


667mod
1008mod
Rdmod
NGOC707mod
NMB1375mod
NMA1590mod
285mod
R2846mod
162mod
1233mod
432mod
1124mod
R2866mod
667mod
1008mod
Rdmod
360
380
400
420
440
460
480
500
520
540
560
580
600
620
640
660
680

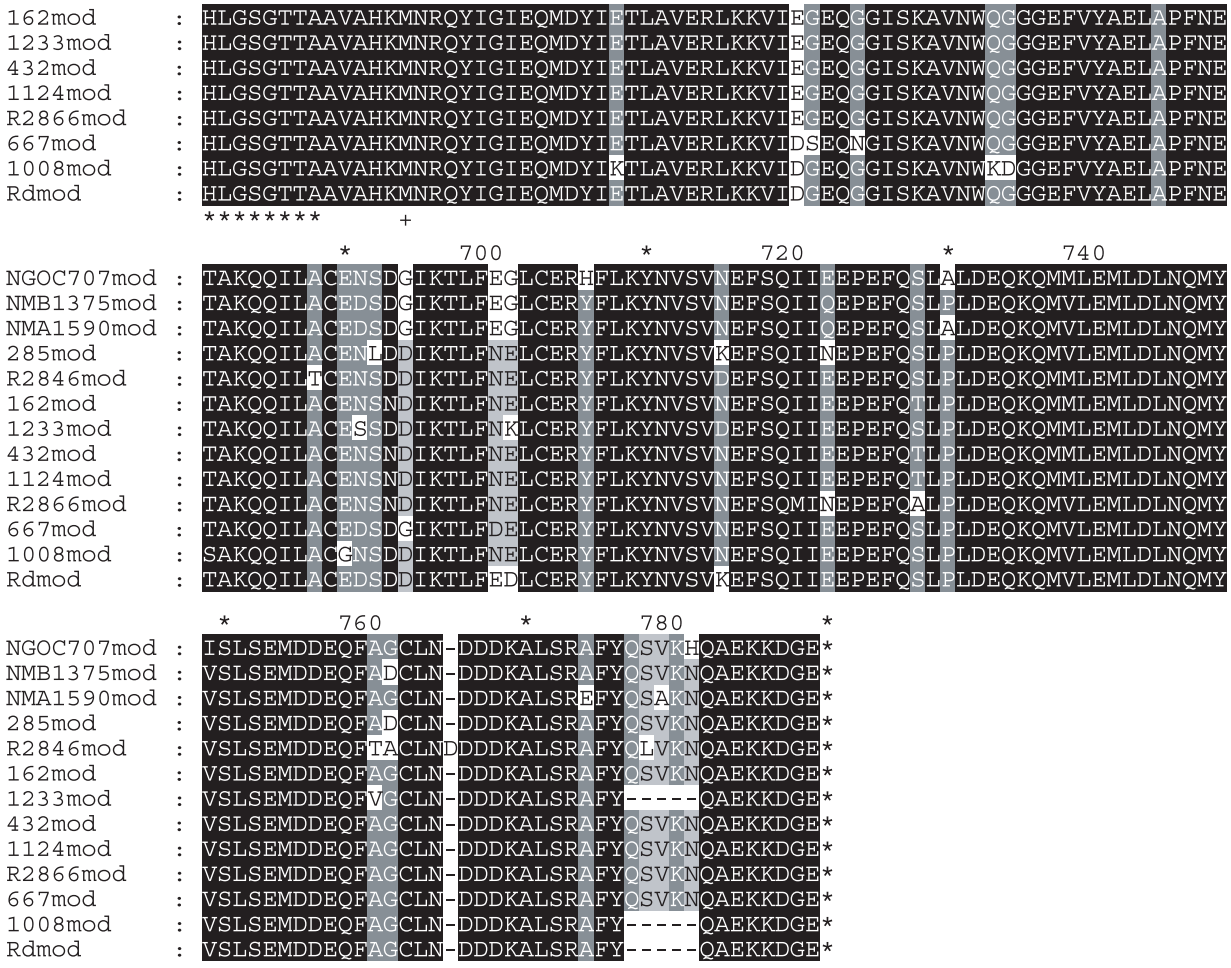


Figure 2. Amino acid sequences of full-length *H.influenzae* Mod proteins. Amino acid sequences were derived from the nucleotide sequences of *mod* for seven NT*Hi* strains. These sequences were aligned with those of related Mod protein sequences present in the genome sequences of *H.influenzae* strains R2866 (R2866mod), R2864 (R2864mod) and Rd (Rdmod), *N.meningitidis* strains MC58 (NMB1375mod) and Z2491 (NMA1590mod) and *N.gonorrhoeae* strain FA1090 (NGOC707mod). The latter four proteins are identical with the following REBASE genes: M.Hindorf1056P, M.NmeBorf1375P, M.NmeAorf1590P and M.NgoorfC707P. Identical and conserved amino acids are highlighted with black or grey backgrounds, respectively. Missing amino acids are indicated by a dot. Plus signs (+) on the bottom line mark semi-conserved motifs present in 69 Type III R-M systems (Supplementary Figure 4A and data not shown). Asterisks (*) mark amino acids that are conserved in 75% of the 69 Mod proteins.

apart from a 1 repeat unit difference in the repeat tract, to those of isolates 1231 and 1232 (Table 1).

The nucleotide (data not shown) and amino acid (Figure 2) sequences of the complete *mod* sequences displayed a hypervariable region between conserved N- and C-terminal regions. The N-terminal domain contained 45 (11%; region iii in Figure 1) and 153 (32%; Regions v and vi in Figure 1) polymorphic sites upstream and downstream of the repeat tract, respectively. The C-terminal domain (region viii in Figure 1) exhibited variation in length, due to different termination codons being utilized in some strains, and contained 115 polymorphic sites (20%). The hypervariable region varied in length from 594 to 750 nt with each sequence type having a variable region of different length and displayed extremely limited homology in both the nucleotide (data not shown) and amino acid sequences (Figure 2). One motif (EYIXVXXK) that is conserved between Type III MTases was found in this region (37) and was partially conserved in all these strains. All other conserved motifs were found in the N- and C-terminal domains of the protein. Structural predictions for these Mod proteins indicated that the variable regions were not structurally conserved, thus this region in the strain 1008 Mod protein consists mainly of alpha-helices whilst the strain 162 Mod protein consists of some alpha-helices and some regions of a strong beta-strand nature (data not shown). This data indicated that the MTase activity of these proteins is encoded by the conserved regions whilst the large size and structural heterogeneity of the hypervariable region suggested that this region encodes a major, alterable function of the Mod protein.

One possible explanation for hypervariability in the central region of the Mod protein was that this domain mediated protein:protein interactions with a domain of similar variability in the Res protein. To examine this theory, complete sequences were obtained for *res* from three isolates and compared to those derived from genome sequences. The genes varied in length from 2771 to 2888 nt and contained 417 polymorphic sites (15%) (Supplementary Figure 4) of which 132 were non-synonymous (Supplementary Figure 5). The *res* gene from strain Rd contained a 2 nt

deletion, not present in the other sequences, that inactivated the gene (and has led to this ORF being annotated as two ORFs in the genome sequence). The absence of significant variability in Res indicated that variability in the central region of Mod was not being driven by a requirement for the maintenance of interactions between the sub-units of the endonuclease.

Analysis of selection in *mod* genes

In order to identify the selection pressures acting on these genes, a ML analysis was performed on the 5′ end of the *H.influenzae mod* gene, using the 22 sequences generated herein and the strain Rd genome sequence. Analysis of the 3′ end of *mod* (see below) was not possible due to the lower number of sequences covering this region. Six models were employed, three of which (M2, M3 and M8) allowed for positive or diversifying selection. Model M2 was significantly favoured over models M0 and M7 but not M1 (data not shown). Models M3 and M8 were the best supported and significantly favoured over the three models not allowing for positive selection (data not shown). Model M3 provided the most sensitive test for positive selection. Under the M3 model, 7.80% of sites were found to experience positive selection pressure ($d_N/d_S = 2.29$); 17.36% of sites were weakly conserved ($d_N/d_S = 0.36$); and the remaining 74.84% of sites were strongly conserved, under intense negative or stabilizing selection pressure ($d_N/d_S = 0.0001$). Under the M8 model, a similar proportion (7.54%) to the M3 model was under positive selection, with a similar d_N/d_S ratio ($d_N/d_S = 2.31$). In the M3 model, posterior probabilities identified eight sites under positive selection above the 90% level (Table 2). The M8 model identified seven of these sites as being under positive selection at the 90% level (site 105 was not identified under M8). This analysis demonstrated the existence of strong selection for diversification of some regions, and therefore certain functions, of these Mod alleles.

Table 2. Positively selected sites in *H.influenzae mod* genes

Codon[a]	Amino acid[b]	d_N/d_S
53*[c]	D	2.206
81*	S	2.256
105	S	2.125
196**	S	2.282
201**	T	2.280
229*	V	2.232
242*	N	2.239
244**	G	2.289

[a]Codon sites are given with reference to the amino acid sequence in Supplementary (Figure 2).
[b]Amino acids are those encoded by codons in the Rd *mod* sequence.
[c]Codons not marked with an asterisk were under positive selection at or above the 90% level; a single asterisk indicates 95% or greater posterior probability of positive selection; and a double asterisk indicates 99% or greater posterior probability.

Phylogenetic relationships of *H.influenzae mod* genes

One strategy for determining the functions and selective pressures acting on this *H.influenzae* Type III R-M system was to examine the phylogenetic relationships between the *H.influenzae mod* gene sequences and other extant Type III MTases. Amino acid sequences of other known and putative Type III MTases were obtained from genome sequence databases and aligned with the amino acid sequences of the full-length *H.influenzae mod* gene sequences. Conserved protein sequences were found in the N- and C-termini of these proteins surrounding a variable region, which could not be aligned (data not shown). The conserved sequences in these proteins exhibited variable degrees of similarity to the ten motifs described by Malone *et al*. (10) with motifs I, II and IV having the highest degree of conservation in the *H.influenzae* Mod proteins (Figure 2) and a beta-organization (III-IV-V-VI-VIII-IX-X-I-II).

To facilitate identification of phylogenetic relationships, the alignments were trimmed to contain only the conserved domains. The NJ algorithm was employed to group similar sequences together (Figure 3A) but was not a reliable indicator of any deeper phylogenetic structure, as indicated by poor bootstrap values on most branches (data not shown). The *H.influenzae* Mod proteins clustered together in a clade that included sequences from the pathogenic *Neisseria*, *Lactococcus lactis*, *M.pulmonis* and *H.pylori*. Notably Mod proteins did not cluster by species. Thus M.NmeBorf1261P clustered with M.EcoPI rather than with M.NmeBorf1375P and the *H.influenzae* Mod proteins. Similarly *H.pylori* Type III Mod proteins were found in three different positions on the tree.

The groupings in this phylogenetic tree permitted a re-evaluation of the alignments of the full-length amino acid sequences. The NT*Hi* and *Neisserial* (i.e. M.NmeB1375P, M.NmeA1590P and M.NgoorfC707P) Mod proteins have an N-terminal region of 202–242 amino acids, which precedes the conserved MTase motifs. This region is of variable length in other Mod proteins. Intriguingly, there was a domain within this region with a high degree of similarity to the N-termini of four other Type III Mod proteins (M.Hpy99orf1296P, M.MpuC orf3980P, M.MpuC orf3960P and M.LlaF1) (data not shown). Homology extended over ~120 amino acids and included a DEI(D/S) motif. Another group of 17 Mod proteins from diverse species also exhibited significant homology to each other in this region [~40 amino acids, motifs of WXGK(K/S) and two conserved prolines]. Finally a conserved domain was present in the N-termini of another group of 9 Mod proteins. This group included the M.StyLTI-like Mods, M.EcoPI, M.EcoP15I and the M.NmeBorf1261-like *Neisserial* Mod proteins [~39 amino acids, motifs (W/F)(I/L)GK(S/D)YA and HNxxENxxS]. The conservation of these motifs between species suggests that the N-terminal region encodes a function that is intrinsic to a particular group of Type III R-M systems and is not a separate specific function of the bacterial species.

The branch with the *H.influenzae* sequences was then re-analysed using amino acids sequences that lacked only the repeat tract and hypervariable regions (Figure 3B). The *H.influenzae* enzymes clustered together with the Neisserial enzymes. A similar cluster of the *Mycoplasma pneumoniae* Mod proteins was also formed. Similar trees were obtained by both NJ and split decomposition analyses (data not shown). Intriguingly M.NmeAorf1590P of *N.meningitidis* serogroup A was most closely related to Mod of *H.influenzae* strain 285 rather than to the *N.meningitidis* serogroup B or *Neisseria gonorrhoeae* Mod proteins.

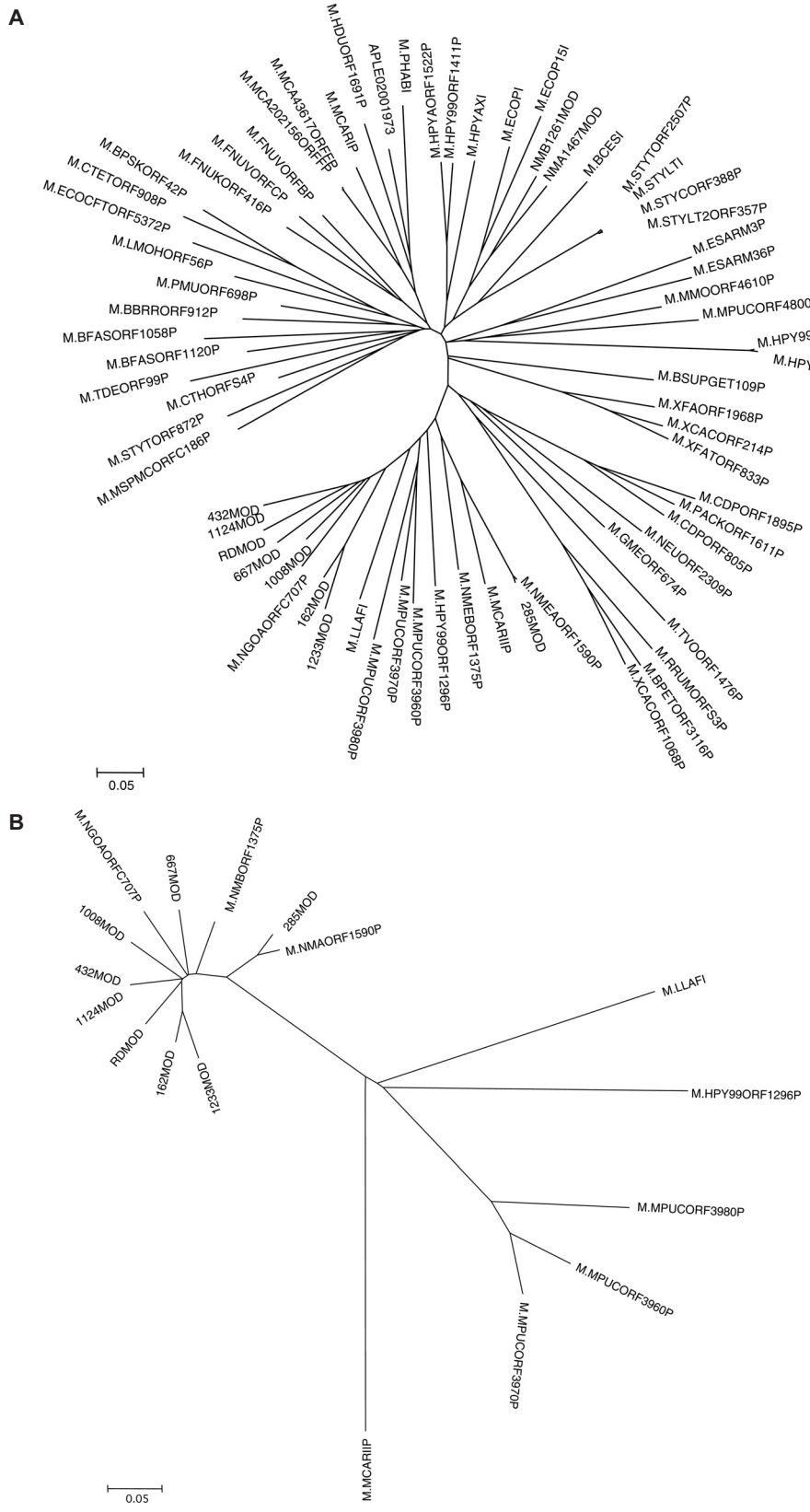


Figure 3. Phylogenetic trees of Mod amino acid sequences from known and putative Type III R-M systems. Amino acids sequences for Mod proteins of Type III R-M systems were obtained from genomic databases and from the REBASE database and aligned with seven full-length NT*Hi mod* amino acid sequences. Alignments were then trimmed manually to include only the semi-conserved regions (**A**) or to exclude the repeat and variable regions (**B**). Alignments included 69 Mod sequences (A) or a sub-set of these sequences (B). Phylogenetic trees were generated using these alignments and a NJ algorithm implemented in the program MEGA, version 3 (34).

Horizontal transfer of *mod* and *res*

The clustering of the *Haemophilus* and Neisserial gene sequences suggested that horizontal gene transfer had occurred and so these sequences were examined in detail. The Neisserial genes all contained a 5′-AGCC repeat tract (Table 1). Alignments of the nucleotide and amino acid sequences with the *H.influenzae* genes indicated significant homology in the regions flanking the repeat tract and within the C-terminus of the protein (data not shown and Figure 2). The hypervariable region was not conserved either between the *H.influenzae* and Neisserial genes or between the Neisserial genes (data not shown) and thus the Neisserial genes defined three further Mod types for this R-M system. The *res* genes were also conserved between these species (data not shown). The Neisserial *mod-res* genes are inserted between a conserved hypothetical open reading frame and a gene encoding L-lactate dehydrogenase. The homology between the Neisserial and *Haemophilus* DNA sequences is limited to the 40 nt upstream of the initiation codon, which includes the potential Shine–Delgarno and promoter elements, and to 10 nt downstream of the *res* termination codon (Supplementary Figure 6). The Neisserial *mod-res* genes exhibited G+C contents of 38–42% close to the average for *H.influenzae* genomes (38%) and significantly different from that of *N.meningitidis* genomes (52–53%), suggesting lateral gene transfer of this Type III R-M system from *H.influenzae* to *N.meningitidis* and *N.gonorrhoeae*.

The horizontal transfer of the *mod* gene between species was an indication of genetic exchange of *mod* gene sequences. This phenomenon was examined further by a statistical analysis of recombination between the NT*Hi mod* genes. Five highly significant recombination fragments were detected using the program GENECONV (data not shown). One of the recombination fragments (339 bp) was in the intergenic region of *mod* types 4 and 8. Two of the fragments, 51 and 89 bp, covered the conserved MTase amino acids of *mod* types 1 and 4 or 2 and 6, respectively. Finally, two of the fragments were found as mosaic sequences in isolates 162, 486 and 1008. Isolates 162 and 486 form *mod* sequence type 5 but isolate 162 has a unique sequence with a number of polymorphisms in the N-terminal region of *mod*. Contrastingly, isolates 486 and 1008 are identical for most of the intergenic region (i.e. from position 11) and the N-terminal region of *mod* but differ in the variable region (1008 has *mod* sequence type 6). The last polymorphism between isolates 162 and 486 is at nucleotide 1047 whilst the first polymorphism between 486 and 1008 is at nucleotide 1116 (Supplementary Figure 1). These results suggested that recombination has occurred between 1047 and 1116 nt in isolate 486 and indicated that the donor DNA was provided by either 162 or 1008 or other isolates with similar DNA sequences.

A statistical analysis of recombination was then performed on alignments of the full-length nucleotide sequences of the *H.influenzae* and Neisserial *mod* genes (data not shown). A total of 63 highly significant recombination fragments were detected. A total of 14 fragments were found in the *N.meningitidis mod* genes and 11 of these were across species. Notably the N-terminal regions of the M.NmeBorf1375P and M.NmeAorf1590P *mod* genes were identical up to the repeat tract and diverged subsequently, indicating that the repeat tract may have acted as a recombination break-point. Subsequent to the repeat tract, *nmeAorf1590MP* exhibited significant similarity to the *mod* gene of *H.influenzae* isolate 285 (90 and 88% at the amino acid and nucleotide levels, respectively), such that these genes were more similar to each other than to genes from the homologous species. Evidence for genetic exchange of *mod* sequences between these strains was provided by the detection of a recombination fragment (75 bp) in the nucleotide sequences of the variable region. This evidence of recombination between *mod* alleles, both within and across species boundaries, provides further support for the existence of strong selective pressures acting on the function(s) of these proteins.

DISCUSSION

A detailed picture of the molecular evolution of a phase variable Type III R-M system from *H.influenzae* is provided in this report. The data covers the degree of allelic diversity (the first of its kind for a Type III R-M system), selection pressures, extent of horizontal gene transfer and variations in the mechanism of PV. Whilst the genes analysed herein have not been shown to encode functional Type III R-M systems, these genes are likely to encode a previously characterized *H.influenzae* Type III R-M system (see Introduction). Thus, these findings are relevant to the identification of the functional domains of Type III R-M systems and to the exploration of the influence of phase variable R-M systems on the evolutionary fitness of bacterial species.

A number of motifs characteristic of MTases of R-M systems were observed in the ends of the *H.influenzae mod* sequences and were conserved in the majority of these sequences (Supplementary Figure 2 and Figure 2). ML analysis indicated that ~75% of sites encoding the N-terminal region of the Mod protein experience strong negative (purifying) selection most likely reflecting a requirement for structural/functional conservation. These results suggest that these *mod* alleles encode functional MTases and that there is a strong selection for the maintenance of this activity in these Mod proteins. In contrast, the central region of these *mod* sequences exhibited significant diversity, much higher than other regions of this locus, such that sequence alignment was not possible for the different Mod types (Figure 2 and Supplementary Figure 3). This diversity, indicative of strong diversifying selection, prevented a statistical analysis of selection. However, the ML analysis identified two sites (196 and 201 amino acids, see Table 2) in the N-terminal region of Mod that were subject to positive diversifying selection. These sites were upstream of the conserved DPPY motif and mark the ends of an insertion, such that different strains encode one of four possible motifs: FDGFLKGV, SDGFLKGV, FDGTP and SDGTP. It is possible that this region is under selection because it is part of the TRD or because modifications in the TRDs require compensatory modifications in this region to maintain MTase activity. The hypervariability of the central regions of the *mod* sequences was not matched by a similar level of diversity in the *res* sequences (Supplementary Figure 4), indicating that the maintenance of interactions between the Mod and Res proteins was not driving diversification

of *mod*. The diversity of these sequences was reminiscent of the plasticity observed for the HsdS sub-units of Type I R-M systems, which confer sequence specific DNA recognition (7). We speculate, therefore, that this region of *mod* encodes a TRD, which is responsible for recognition of specific DNA sequences by the Mod protein and is under strong diversifying selection.

Horizontal transfer of R-M systems is an established phenomenon (11,12,38). One impact of such transfer events could be alterations to the recognition site of an R-M system. The observation of a mosaic sequence in the *mod* genes of NT*Hi* isolates 162, 486 and 1008 provided a clear example of genetic exchange between alleles of this *H.influenzae* R-M system. The break-point for this recombination event was adjacent to the variable region and may, depending on the direction of transfer, have resulted in exchange of an entire variable region between two *H.influenzae* strains and hence re-assortment of the putative TRD domains. Another potential example of transfer of a variable region was provided by analysis of the *mod* genes of NT*Hi* isolate 285 and *N.meningitidis* strain Z2491 (i.e. M.NmeAorf1590P). These phylogenetically-related *mod* alleles exhibited significant similarity from the repeat tract to the end of the gene (see Figure 3B) and shared a recombination fragment of 75 bp. One interpretation of this data is, that the genetic exchange occurred between these two strains (or strains containing related *mod* sequences) and involved a large fragment which then diverged by mutation. This large fragment included the variable region and hence the putative TRD. Intriguingly, similarity between these *mod* alleles was found at either end of the variable region (43/44 and 64/66 identical amino acids in regions 350–394 and 523–604, respectively, Figure 2) whilst the intervening sequences were divergent. This arrangement is indicative of a further recombination event within the variable region of one of these *mod* alleles and may be evidence of transfer of a TRD sub-domain encoded within the variable region. The high nucleotide sequence conservation of these and other *mod* alleles derived from *H.influenzae* and Neisserial isolates, the conservation of the flanking regions of the *mod-res* locus (Supplementary Figure 6), and the low G + C% of the Neisserial sequences are indicative of recent transfer of this Type III R-M system from *H.influenzae* into the Neisserial spp. Indeed, it is possible that transfer of these genes or fragments of these genes has occurred multiple times. An implication of this transfer and of the hypervariability of the Neisserial *mod* sequences is that similar selective pressures are acting on this system in all these species. Confirmation of the extent of horizontal transfer of *mod* sequences within and between these species and of the affects of these transfer events on the functions of this R-M system will require a combination of further molecular genetic analyses and functional studies.

PV of *mod* is mediated by a tetranucleotide repeat tract (1). As reported previously the repeat tract is absent in a number of the NT*Hi* strains and of variable length in other strains (1). The assignation of Mod types to these sequences permits a further evaluation of the variability of the repeat tracts. Nine of the thirteen Mod types are associated with repeats and in all the types for which multiple isolates were identified the repeat tract was of variable length (Table 1). Only in Mod type 2 is the same reading frame maintained in all the isolates. These isolates are likely to be ON phase variants as the frame is initiated by the distal 5′-ATG, which is associated with high expression. Seven of the isolates have repeat numbers associated with the proximal initiation codon or no in-frame initiation codon. The level of expression from the proximal initiation codon has been reported to generate undetectable or low levels of gene expression in *lacZ* reporter constructs (1,20), thus it is likely that these isolates are OFF phase variants.

These observations of significant variability in the repeat tract combined with the diversity in the central region of *mod*, especially if the hypothesis of an impact on sequence recognition is accepted; impinge on the possible roles of this phase variable Type III R-M system. One proposal is that the MTases of this R-M system mediate stochastic variations in expression of a transcriptional regulon as a mechanism for adaptation to environmental fluctuation (20). This hypothesis is based on the observation that PV of the *mod* gene of *H.influenzae* strain Rd (one of the genes analysed in this paper) is linked to phase variable expression of other non-phase variable genes and is predicated on the notion that the Mod protein controls expression of specific genes through methylation of promoter elements (20). This phenomenon could only arise with strong negative selection, resulting in maintenance of the recognition sequence of the Mod protein, such that significant variability in the *mod* gene would not have been expected. In particular, variations in the *mod* recognition sequence, as proposed herein, would have prevented evolution of specific promoters subject to control by Mod MTase activity. Furthermore, alterations in the putative TRD due to horizontal transfer, as described above, would unlink *mod* alleles from their associated promoters. An alternative possibility is that this R-M system acts as a barrier to genetic exchange through transformation and that PV permits temporary alleviation of this barrier and occasional acquisition of beneficial sequences, as suggested for the R-M systems of other bacterial species (17–19). In this case, the variability in the putative TRD suggests that there is strong selection for sexual isolation of *H.influenzae* clones, a suggestion which contrasts with the high rates of recombination observed for many other genes within this same collection of *H.influenzae* isolates (36).

A third proposal is that this phase variable R-M system has a role in preventing bacteriophage infections. It has been speculated that the evolution of diversity in bacteriophage genomes provides resistance to R-M systems by altering the number of recognition sequences and that this diversity acts as a selection pressure for evolution of new target recognition specificities in the R-M system (7,39,40). Three possible explanations may then be invoked for evolution of phase variable resistance to bacteriophage infection. First, PV may limit generation of resistant bacteriophages. Selection for expression of an R-M system (i.e. ON variants) is likely to weaken as the resistance of the bacteriophage develops resulting in an increase in the number of OFF phase variants in the bacterial population. Oscillation between replication of the phage in these ON and OFF variants may prevent evolution of fully resistant phage. Second, methylation of some 'self' recognition sequences may be detrimental for bacterial proliferation, such that PV may permit switching

between a bacteriophage resistant but inferior growth phenotype and a bacteriophage sensitive but superior growth phenotype. This latter system would permit deterimental MTase recognition sequences to be carried in the bacterial population (i.e. in the OFF state) until infection by a bacteriophage infection provides selection for expression of this allele. Third, PV may permit temporary loss of the phage resistance phenotype and incorporation of foreign, potentially beneficial DNA by transduction [as proposed by Zaleski *et al*. (24) for HindI]. Whilst it is unclear if infrequent acquisition of adaptive phenotypes by transduction would exert a high enough selection for evolution of phase variable R-M systems, this proposal may have important implications for the evolutionary fitness of *N.meningitidis*. An active prophage was recently found associated with disease causing isolates of this species (41) and it has been argued that this prophage confers a fitness advantage by increasing transmission of the meningococcus between hosts (42). It is highly likely that transfer of this prophage between strains will be subject to control by R-M systems (including the system described herein), such that switching off an R-M system by PV would provide an opportunity for meningococci not carrying the prophage to acquire it thereby increasing their fitness.

ACKNOWLEDGEMENTS

The authors thank Alison Cody and Gaynor Jenkins for help with preparation of genomic DNA and for electrophoretic separation of sequencing reaction products. The authors are also grateful to Derek Hood for his suggestions and Dlawer Ala'Aldeen for support during writing up of this project. C.D.B. was supported by a Wellcome Trust programme grant, 070123/Z/02/Z entitled 'Mutation rates of simple sequences and their contribution to the pathogenesis of *H.influenzae* and *N.meningitidis*' and a Value in People Award from the University of Nottingham. Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. De Bolle,X., Bayliss,C.D., Field,D., van de Ven,T., Saunders,N.J., Hood,D.W. and Moxon,E.R. (2000) The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol. Microbiol.*, **35**, 211–222.
2. 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.
3. Dybvig,K. and Yu,H. (1994) Regulation of a restriction and modification system via DNA inversion in *Mycoplasma pulmonis*. *Mol. Microbiol.*, **12**, 547–560.
4. Ryan,K.A. and Lo,R.Y. (1999) Characterization of a CACAG pentanucleotide repeat in *Pasteurella haemolytica* and its possible role in modulation of a novel type III restriction-modification system. *Nucleic Acids Res.*, **27**, 1505–1511.
5. Saunders,N.J., Jeffries,A.C., Peden,J.F., Hood,D.W., Tettelin,H., Rappouli,R. and Moxon,E.R. (2000) Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Micro.*, **37**, 207–215.
6. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S., Dryden,D.T., Dybvig,K. *et al*. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
7. Murray,N.E. (2000) Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Bio. Rev.*, **64**, 412–434.
8. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.
9. Bourniquel,A.A. and Bickle,T.A. (2002) Complex restriction enzymes: NTP-driven molecular motors. *Biochimie*, **84**, 1047–1059.
10. 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.
11. Bujnicki,J.M. (2001) Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons. *Acta Biochim. Pol.*, **48**, 935–967.
12. Nobusato,A., Uchiyama,I. and Kobayashi,I. (2000) Diversity of restriction-modification gene homologues in *Helicobacter pylori*. *Gene*, **259**, 89–98.
13. Dybvig,K., Sitaraman,R. and French,C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 13923–13928.
14. Kobayashi,I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.*, **29**, 3742–3756.
15. Murray,N.E. (2002) 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology*, **148**, 3–20.
16. van der Woude,M.W. and Baumler,A.J. (2004) Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.*, **17**, 581–611.
17. Donahue,J.P., Israel,D.A., Peek,R.M., Blaser,M.J. and Miller,G.G. (2000) Overcoming the restriction barrier to plasmid transformation of Helicobacter pylori. *Mol. Microbiol.*, **37**, 1066–1074.
18. Ando,T., Xu,Q., Torres,M., Kusugami,K., Israel,D.A. and Blaser,M.J. (2000) Restriction-modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. *Mol. Micro.*, **37**, 1052–1065.
19. Seib,K.L., Peak,I.R.A. and Jennings,M.P. (2002) Phase variable restriction-modification systems in *Moraxella catarrhalis*. *FEMS Immunol. Med. Micro.*, **32**, 159–165.
20. 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.
21. Gumulak-Smith,J., Teachman,A., Tu,A.H., Simecka,J.W., Lindsey,J.R. and Dybvig,K. (2001) Variations in the surface proteins and restriction enzyme systems of *Mycoplasma pulmonis* in the respiratory tract of infected rats. *Mol. Microbiol.*, **40**, 1037–1044.
22. Hood,D.W., Deadman,M.E., Jennings,M.P., Bisercic,M., Fleischmann,R.D., Venter,J.C. and Moxon,E.R. (1996) DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. USA*, **93**, 11121–11125.
23. Moxon,E.R., Rainey,P.B., Nowak,M.A. and Lenski,R. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, **4**, 24–33.
24. Zaleski,P., Wojciechowski,M. and Piekarowicz,A. (2005) The role of Dam methylation in phase variation of *Haemophilus influenzae* genes involved in defence against phage infection. *Microbiology*, **151**, 3361–3369.
25. Kelly,T.J.,Jr and Smith,H.O. (1970) A restriction enzyme from *Hemophilus influenzae II*. *J. Mol. Biol.*, **51**, 393–409.
26. van Belkum,A., Scherer,S., van Leeuwen,W., Willemse,D., van Alphen,L. and Verbrugh,H.A. (1997) Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect. Immun.*, **65**, 5017–5027.
27. Piekarowicz,A. (1982) HineI is an isoschizomer of HinfIII restriction endonuclease. *J. Mol. Biol.*, **157**, 373–381.
28. Piekarowicz,A. and Brzezinski,R. (1980) Cleavage and methylation of DNA by the restriction endonuclease HinfIII isolated from *Haemophilus influenzae* Rf. *J. Mol. Biol.*, **144**, 415–429.

29. Piekarowicz,A., Bickle,T.A., Shepherd,J.C.W. and Ineichen,K. (1981) The DNA sequence recognised by the HinfIII restriction endonuclease. *J. Mol. Biol*., **146**, 167–172.
30. Nicholas,K.B., Nicholas,H.B.,Jr and Deerfield,D.W.II (1997) GeneDoc: Analysis and Visualisation of Genetic Variation. EMBNE. NEWS 4:14.
31. Swofford,D. (1998) *PAUP* Phylogenetic Anaysis Using Parsimony and Other Methods*. Sinauer, Sunderland, Mass, USA.
32. Woelk,C.H. and Holmes,E.C. (2001) Variable immune-driven natural selection in the attachment (G) glycoprotein of respiratory syncytial virus (RSV). *J. Mol. Evol*., **52**, 182–192.
33. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.
34. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*., **5**, 150–163.
35. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2005) REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*., **33**, D230–D232.
36. Cody,A.J., Field,D., Feil,E.J., Stringer,S., Deadman,M.E., Tsolaki,A.G., Gratz,B., Bouchet,V., Goldstein,R., Hood,D.W. *et al*. (2003) High rates of recombination in otitis media isolates of non-typeable *Haemophilus influenzae*. *Infect. Gen. Evolution*, **3**, 57–66.
37. Dartois,V., De Backer,O. and Colson,C. (1993) Sequence of the *Salmonella typhimurium* StyLT1 restriction-modification genes: homologies with EcoP1 and EcoP15 type-III R-M systems and presence of helicase domains. *Gene*, **127**, 105–110.
38. 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.
39. Tock,M.R. and Dryden,D.T. (2005) The biology of restriction and anti-restriction. *Curr. Opin. Microbiol*., **8**, 466–472.
40. Bickle,T.A. and Kruger,D.H. (1993) Biology of DNA restriction. *Microbiol. Rev*., **57**, 434–450.
41. Bille,E., Zahar,J.R., Perrin,A., Morelle,S., Kriz,P., Jolley,K.A., Maiden,M.C., Dervin,C., Nassif,X. and Tinsley,C.R. (2005) A chromosomally integrated bacteriophage in invasive meningococci. *J. Exp. Med*., **201**, 1905–1913.
42. Moxon,E.R. and Jansen,V.A. (2005) Phage variation: understanding the behaviour of an accidental pathogen. *Trends Microbiol*., **13**, 563–565.