ModM DNA methyltransferase methylome analysis reveals a potential role for *Moraxella catarrhalis* phasevarions in otitis media

Luke V. Blakeway[1], Peter M. Power[1], Freda E.-C. Jen[1], Sam R. Worboys[1], Matthew Boitano[2], Tyson A. Clark[2], Jonas Korlach[2], Lauren O. Bakaletz[3], Michael P. Jennings[1], Ian R. Peak[1,4^], Kate L. Seib[1^#]

[1]Institute for Glycomics, Griffith University, Southport, Queensland, Australia
[2]Pacific Biosciences, Menlo Park, CA, USA
[3]Center for Microbial Pathogenesis, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio, USA
[4]School of Medical Science, Griffith University, Southport, Queensland, Australia
^These authors equally conceived, designed, and directed this study.
[#]Corresponding author. Mailing address: Institute for Glycomics, Griffith University, Parklands Dr., Southport, Queensland, Australia. Tel., +61 755527453. E-mail address, k.seib@griffith.edu.au.

Running title: *Moraxella catarrhalis ModM phasevarion*

Nonstandard abbreviations

Mod; DNA methyltransferase

SMRT; single molecule, real-time

iTRAQ; Isobaric tags for relative and absolute quantitation

Abstract

Moraxella catarrhalis is a significant cause of otitis media and exacerbations of chronic obstructive pulmonary disease. Here we characterize a phase variable DNA methyltransferase (ModM), which contains 5'-CAAC-3' repeats in its open reading frame that mediate high frequency mutation resulting in reversible ON/OFF switching of ModM expression. Three *modM* alleles have been identified (*modM1-3)* with *modM2* being the most commonly found allele. Using single molecule, real-time (SMRT) genome sequencing and methylome analysis, we have determined that the ModM2 methylation target is 5'-GAR^{m6}AC-3', and 100% of these sites are methylated in the genome of the *M. catarrhalis* 25239 ModM2 ON strain. Proteomic analysis of ModM2 ON and OFF variants revealed that ModM2 regulates expression of multiple genes that have potential roles in colonization, infection and protection against host defenses. Investigation of the distribution of *modM* alleles in a panel of *M. catarrhalis* strains, isolated from the nasopharynx of healthy children or middle ear effusions from patients with otitis media, revealed a statistically significant association of *modM3* with otitis media isolates. The modulation of gene expression via the ModM phasevarion (phase variable regulon), and the significant association of the *modM3* allele with otitis media, suggests a key role for ModM phasevarions in the pathogenesis of this organism.

Keywords:

epigenetic regulation; middle ear infection; phase variation; single molecule, real-time (SMRT) genome sequencing

Introduction

Moraxella catarrhalis is a Gram-negative diplococcus and a human-restricted respiratory tract pathogen. It is the third most common cause of childhood otitis media (OM), after non-typeable *Haemophilus influenzae* and *Streptococcus pneumoniae*, being associated with approximately 25% of acute OM cases and 46% of chronic middle ear effusions (1-3). *M. catarrhalis* is also commonly isolated from the upper respiratory tract of apparently asymptomatic children, with more than 70% of children colonized at least once by the age of 1, and the frequency of colonization being associated with development of OM (4-7). OM is the most common bacterial infectious disease in childhood, and is the main cause of pediatric antibiotic usage and visits to healthcare providers in the developed world (8-10). *M. catarrhalis* is the second most common cause of exacerbations of chronic obstructive pulmonary disease (COPD) after *H. influenzae*, being responsible for 10-15% of exacerbations (~2-4 million episodes per year in the United States) (11, 12). COPD is a chronic debilitating disease that is one of the top five causes of death worldwide (13).

Due to the importance of *M. catarrhalis* as a human pathogen and the finding that the frequency of colonization with *M. catarrhalis* is rising in certain regions where pneumococcal conjugate vaccines are widely used (14), it is increasingly important to understand its mechanisms of pathogenesis. Phase variation, the high frequency ON/OFF or graded switching of gene expression, is a common feature of host adapted bacterial pathogens such as *M. catarrhalis, H. influenzae* and *Neisseria meningitidis,* which enables them to adapt to changing host microenvironments and evade host immune responses (15). Phase variation is typically mediated by simple DNA repeats, within the open reading frame (ORF) of a gene, that exhibit reversible loss/gain of repeat units during replication, resulting in frame shift mutations. We have reported that phase variation of N^6-adenosine DNA methyltransferases (Mod) in several pathogens results in global changes in gene expression coincident with variations in methylation of the genome in the Mod ON and OFF states (reviewed in 16). These phase variable regulons, known as phasevarions, have been studied in *H. influenzae* (ModA) (17, 18), the pathogenic *Neisseria* (ModA, ModB, ModD) (19, 20) and *Helicobacter pylori* (ModC)(21). In these pathogens, phase variable Mods control expression of surface antigens and virulence factors, leading to altered phenotypes between Mod ON and OFF variants. The ModA-D proteins have different alleles based on amino acid differences in their central DNA recognition domain (19, 20, 22), and each allele methylates different DNA sequences (23). DNA methylation mediates epigenetic regulation by altering the interaction of DNA-binding regulatory proteins with their cognate DNA sequence (24).

We have previously identified two potentially phase variable DNA methyltransferases in *M. catarrhalis* (25). Here we describe the distribution of *modM1-3* alleles, identify the ModM2 DNA methylation sequence using single molecule, real-time (SMRT) methylome analysis,

characterize the ModM2 phasevarion using proteomic analysis, and examine potential for association of the *modM* alleles with OM disease.

Materials and Methods

Bacterial strains and growth conditions

M. catarrhalis strains used in this study include ATCC 8193, ATCC 25239, ATCC 23246, CCUG 26404, CCUG 3292, 4223, 26297 and 17 middle ear isolates from OM-prone children, 9 nasopharyngeal isolates from children with OM, and 55 nasopharyngeal isolates from "normal healthy children" (supplied by Dr. Lauren O. Bakaletz (Centre for Microbial Pathogenesis, The Research Institute at Nationwide Children's, Columbus, Ohio, USA)). For some patients, samples were obtained from two distinct anatomical sites; the middle ear and nasopharynx of the same patient. *M. catarrhalis* strains were grown on Brain Heart Infusion Agar (BHI, Oxoid, Hampshire, UK) at 37°C with 5% CO_2, with kanamycin (20 μg/ml) as required. *Escherichia coli* strains were cultured with Luria-Bertani medium (Difco, Franklin Lakes, NJ, USA), with ampicillin (100 μg/ml) or kanamycin (20 μg/ml) as required.

Generation of *modM* mutant strain

To generate the *modM2* knockout mutant of 25239, the pMcrepI plasmid (25) was subjected to inverse PCR with primers SWF (5'-TACCCGGGCAATACTGATAAATATCACGCCATC-3'; Restriction enzyme sites added for cloning purposes are underlined) and SWG (5'-TAGGATCCGCAGGTTCGGCAACCACC-3') to remove the central DNA recognition domain and to introduce a BamHI restriction enzyme site. The kanamycin-resistance gene was excised from pUC4K (Pharmacia, Piscataway, NJ, USA) using BamHI (New England Biolabs, (NEB), Ipswich, MA, USA) and inserted into the BamHI site in *modM*. The pMcrepI::*kan* plasmid was linearized with BsaI and used to transform 25239, as previously described (26).

Analysis of ModM expression

The length of the 5'-CAAC-3' repeat tract in *modM*, and the percentage of each fragment length, were determined by GeneScan fragment analysis (Applied Biosystems, Grand Island, NY, USA) (19, 20). The primer pair Mcmod2F-6FAM (5'-[6FAM]TTACTTGACACTCTGAATGGA-3') and McmodrepR2 (5-GTATTATGGGCAGTTTTTAAG-3') was used to amplify the repeat region of *modM2*. 25239 *modM2* ON and OFF colonies were isolated and quantified using GeneScan fragment length analysis as previously described (20). The strains described as ModM ON and

ModM OFF in all experiments had >90% homogeneity with respect to the ON/OFF status of ModM2.

Western blot analysis of ModM expression in whole cell lysates of *M. catarrhalis* strains was performed as previously described (27). Cell lysates were prepared after overnight growth on agar plates, and 10 µl (equivalent to a final optimal density at 600nm of 5) loaded on a 4-12% Bis-Tris NuPAGE® polyacrylamide gel. A polyclonal anti-Mod antibody (20) that cross reacts with ModM was used.

Distribution analysis and typing of *modM* alleles

Basic Local Alignment Search Tool (blastn and blastp; http://blast.ncbi.nlm.nih.gov/Blast.cgi) searches were used to identify and characterize *mod* alleles in available *M. catarrhalis* genomes, using the *modM1 and modN1* sequences of *M. catarrhalis* strain 23246 (GenBank Accession AY049056 and AY049057, respectively). The *modM* alleles were aligned using ClustalW.
For subsequent typing of *modM* alleles from *M. catarrhalis* isolates, the *modM* DNA recognition domain was amplified by PCR using primers SWD (5'-GATGGCGTGATATTTATCAGTATTG-3') and SWE (5'-GGTGGTTGCCGAACCTGC-3'). PCR products were digested separately with the restriction endonucleases AluI, AseI and EcoRI (NEB) and the resulting digest products were resolved by electrophoresis on a 2% agarose gel.

Methylome determination by SMRT

The 23259 *modM2* ON (21 repeats), OFF (20 repeats) and the *modM2::kan* mutant strains were grown overnight on BHI agar, and genomic DNA extracted using the QIAGEN (Hilden, Germany) Genomic-tip 20/G kit as per manufacturer's instructions. SMRTbell libraries were prepared as previously described (28). Briefly, genomic DNA was sheared to an average length of 1.5 kb using adaptive focused acoustics (Covaris, Woburn, MA, USA) for short insert libraries, and approximately 10 kb using g-TUBEs (Covaris) for long insert libraries. Libraries were then damaged repaired, end repaired, and ligated to hairpin adapters. Incompletely formed SMRTbell templates were digested using Exonuclease III (NEB) and Exonuclease VII (Affymetrix, Cleveland, OH, USA). Sequencing was carried out on the *PacBio RS II* (Menlo Park, CA, USA) using standard protocols for short and long insert libraries.

Bioinformatic analysis

The whole genome sequence determined by SMRT were annotated using the Prokka bacterial genome annotation tools and the prepared file submitted to GenBank (accession number CP007669). *modM* alleles were aligned using ClustalW and visualized with JalView, and

phylogenetic reconstruction was performed using a neighbor-joining uncorrected ("p") distance matrix.

Restriction inhibition assays

Genomic DNA was extracted from the 23259 *modM2* ON (21 repeats), OFF (20 repeats) and the *modM2::kan* mutant strains and digested for 1hr with BsmBI or HindIII (NEB), according to the manufacturer's instructions.

iTRAQ analysis

M. catarrhalis strain 25239 ModM2 ON and ModM2 OFF were grown on BHI agar for 4 hr. Bacteria were harvested in PBS and duplicate plates pooled, washed, and bacteria were lysed by 4 cycles of freezing at -80^{o}C and thawing on ice. 1D Isobaric tags for relative and absolute quantitation (iTRAQ) was performed on the whole cell lysates of 2 biological replicates (of pooled duplicates) of ModM2 ON and OFF by Australian Proteome Analysis Facility (APAF, Melbourne, VIC, Australia) as previously described (29). For protein preparation, 0.1% SDS was added to cell lysates, and sonic probed on ice (2 cycles, 30 sec). Samples were then buffer exchanged into 0.25 M triethyl ammonium bicarbonate (TEAB), 0.05% SDS (using Viva2, 5 kDa spin filters). Protein concentrations were determined using Direct Detect system (Merck-Millipore, Darmstadt, Germany). A total of 100 µg of protein was reduced with 5 mM Tris (2-carboxyethyl) phosphine (TCEP) for 1 hr at 60 °C, alkylated with 10 mM *s*-methylmethanethiosulfonate (MMTS) at room temperature for 10 min and digested overnight with trypsin at 37°C (1:25 ratio, trypsin to protein). Digested samples were labeled with the iTRAQ reagents following the protocol provided by the vendor (AB Sciex, Foster City, CA, USA). Briefly, one vial of iTRAQ labeling reagent was used for every 100 µg of protein. The iTRAQ reagent was solubilized with ethanol then added to the peptide sample. After 1 h of iTRAQ labeling, the reaction was quenched by adding 70 µL of water. The samples were then mixed at equal ratios (10 µL each) and dried by centrifugal evaporation. The iTRAQ labelled pooled sample was re-suspended in 25 mM TEAB, passed through a detergent removal spin column (Thermo Fisher Scientific, Waltham, MA USA) and then dried by centrifugal evaporation.

Mass Spectrometry (MS) analysis

In-gel trypsin digestion was performed based published methods (30). The coomassie-stained band was excised and destained by washing with 100 mM ammonium bicarbonate/acetonitrile (ACN) (1:1). The in-gel protein was reduced and alkylated by 50 mM dithiothreitol and 100 mM iodoacetamide, respectively. Trypsin was then added to the gel for digestion at 37 °C for 16 hr. In-

gel trypsin digested peptides were finally extracted out from the gel by sonicating in 50% ACN/5% TFA buffer. The trypsin digested sample was concentrated in a Freeze Dryer to remove ACN, and resuspended in 5% ACN/0.1% TFA to 20 µl total volume and run on Shimadzu Prominence nanoHPLC system and ABSCIEX 5600 Triple TOF Mass Spectrometer. Data was processed and analyzed by *ProteinPilot*™ Software.

Results

Phase variable DNA methyltransferases in *M. catarrhalis*

Two putative phase variable DNA methyltransferases were previously identified in *M. catarrhalis* ATCC 23246 (25) and have been renamed as ModM1 (M.McaM1) and ModN1 (M.McaN1), in keeping with nomenclature for DNA methyltransferases and phasevarions (16, 23) (renamed from M.McaRI and M.McaRII (GenBank Accession AY049056 and AY049057, respectively). The *modM1* and *modN1* genes contain a 5'-CAAC-3' repeat tract in the 5' end of their ORFs (Fig. 1A), which is predicted to mediate phase variation. These genes share 46% identity over their coding sequence. The *modM1* gene is predicted to encode a protein of 636 amino acids (MW 72.7 kDa). However *modN1* in ATCC 23246 contains a poly A tract that is the site of a frameshift resulting in early termination of translation. Bioinformatic analysis of these methyltransferase genes in 13 available genomes (Table 1) as well as analysis of 7 other strains indicated that *modM* is present in all strains while the *modN* gene is only present in ATCC 23246, the strain originally examined (25) (based on >90% identity to *modM1* or *modN1* along the length of the deduced amino acid sequence, excluding the DNA recognition domain). Sequence analysis of the putative DNA-recognition domain of the *modM* gene revealed the presence of 3 different alleles (*modM1-3*; Fig. 1A-C). Strain ATCC 23246 is unique in being the only strain from 20 investigated to contain *modM1*, while 17 strains contain *modM2*, and 2 strains contain *modM3*. ModM1–3 share 91-93% identity and 93-96% similarity along the length of the deduced amino acid sequence, excluding the central variable DNA-recognition domain, which shares only 13-31% identity and 30-48% similarity between the alleles (Fig. 1B-C). A strain was defined as having the *modM1*, *modM2* or *modM3* allele if the DNA recognition domain was ≥95% identical at the nucleotide level to the *modM* gene of *M. catarrhalis* strain 23246, 25239 or BC1, respectively (see Table 1 and Figure 1 for details). Figure 1D shows the clustering of the *modM* alleles into three groups.

The number of 5'-CAAC-3' repeat units present ranged from 13 to 34 between strains, suggesting that the *modM* gene has phase variable expression. In addition, the region upstream of the tetranucleotide CAAC repeats contains homopolymeric A and T tracts, which also vary in length between strains, as well as a 16bp sequence that is directly repeated in some strains (Fig.

2A). The *modM* gene has potential initiation codons in two frames, and the homopolymeric tracts may also mediate phase variable expression of ModM at the transcriptional and/or translation level depending on which ATG is the actual start site.

M. catarrhalis strain 25239 was selected for further analysis of the predominant *modM* allele, *modM2*. Due to the complexity of the multiple repeats, it was impossible to predict the actual ATG *in silico*, as expression could initiate from either frame 1 or 3 (one ATG in frame 1 (distal) and two ATGs in frame 3 (proximal); Fig. 2A-B) depending on the number of different repeat units present. Frame 2 (OFF) has no candidate ATG upstream of the 5'-CAAC-3' repeats. In order to investigate ModM expression in the three frames, single colonies of *M. catarrhalis* strain 25239 were screened by DNA sequencing and GeneScan fragment length analysis to isolate *modM2* with 19, 20 or 21 repeat tracts. During this screening, *modM2* repeat tracts ranging from 18-22 were observed, demonstrating phase variation of *modM2* (Fig. 3A). The length of the repeats, and the percentage of the population containing each length, in each sample were determined by GeneScan fragment length analysis using primers Mcmod2F-6Fam/McmodR2 (Fig. 3A). Western blot analysis of samples with a high proportion (>90%) of each repeat length confirmed the expression of ModM2 in the sample containing 21 repeats and the absence of detectable expression in the samples containing 19 or 20 repeats (Fig. 3B). Translation therefore initiates from one of the ATGs proximal to the 5'-CAAC-3' repeats (Fig. 2 and 3C). An isogenic mutant that does not express the *modM2* gene was generated by insertion of a kanamycin resistance cassette into the open reading frame of the gene and the absence of ModM2 was confirmed by Western blot analysis (Fig. 3B).

SMRT methylome analysis: ModM2 methylates 5'-GAR^{m6}AC-3'

M. catarrhalis 25239 *modM2* ON (21 repeats), *modM2* OFF (20 repeats) and *mod::kan* knockout strains were subjected to SMRT genome sequencing, assembly and methylome analysis (28). This process involved complete *de novo* sequencing of strain 25239, and the closed genome sequence is deposited in GenBank (accession number CP007669). The 6-methyladenine (^{m6}A) and 5-methylcytosine (^{m5}C) bases in the genomes were identified by their kinetic signatures, and these were then aligned and clustered to identify the consensus recognition sequences for the active methyltransferases in these strains. Three distinct methylation patterns were detected in 25239 *modM2* ON; 5'-GAR**^{m6}A**C-3', 5'-AC**^{m6}A**N$_6$CTTG-3' (where T represents **^{m6}A** methylation on the opposite strand) and 5'-GCNG**^{m5}C**-3' (where G represents **^{m5}C** methylation on the opposite strand) (Fig. 4).

From comparison of the *modM2* ON to the *modM2* OFF and the *modM2::kan* strains, it is clear that ModM2 is active and forms ^{m6}A residues in the sequence 5'-GAR**^{m6}A**C-3' (Fig. 4A, Fig. S1). 100% of the 3321 5'-GAR**A**C-3' sites present in the genome were detected as methylated, with

35.6% (1183) of these sites being 5'-GAGAC-3' and 64.4% (2138) being 5'-GAAAC-3'. To confirm the ModM2 DNA methylation site, genomic DNA isolated from *modM2* ON, *modM2* OFF and *modM2::kan* was digested with the methylation-sensitive restriction enzyme BsmBI. Numerous BsmBI sites in the genome overlap with the 5'-GAR^{m6}**A**C-3' ModM2 methylation site (Fig. 4B) and DNA digestion by BsmBI is inhibited in the *modM2* ON strain (Fig. 3C). The methylation-insensitive enzyme HindIII was used as a control and is able to digest all three samples.

All of these data and evidence supporting the identification of the ModM2 specificities have been reported to REBASE (31). The methyltransferases responsible for methylating the other sequences in the genome have not been experimentally determined. However, the 5'-GCNG^{m5}**C**-3' site has been assigned to the DNA methyltransferase encoded by 25239_00017, based on 100% sequence homology to M.McaRHORF1876P from *M. catarrhalis* strain BBH18 that is annotated in REBASE as predicted to methylate this site (REBASE Enzyme Number 25867).

The ModM2 Phasevarion; analysis of differentially expressed proteins in *modM2* ON and *modM2* OFF variants

To determine whether phase-variable expression of *modM2* leads to alteration in global gene expression via epigenetic regulation, *M. catarrhalis modM2* ON and *modM2* OFF variants were compared by 1D iTRAQ proteomic analysis. A total of 34 genes had an expression ratio of ≥1.3 fold between the ModM2 ON and OFF variants (p-value ≤ 0.05). Specifically, 19 genes had increased expression in M0579 ModM2 ON, while 15 genes had decreased expression (Table 2), confirming that ModM2 phase variation influences expression of multiple genes, consistent with its definition as a phasevarion (17). The ModM2 phasevarion includes proteins involved in colonization, infection, resistance to host defenses, or other experimental conditions that may mimic those encountered in a human host (see discussion for details).

***modM* allele distribution; *modM3* is associated with middle ear isolates**

To investigate the distribution of *modM* alleles in a large panel of *M. catarrhalis* strains, PCR and restriction digest analysis of the DNA recognition domain was performed. Overall, 81 *M. catarrhalis* isolates were examined, consisting of 64 nasopharyngeal isolates and 17 middle ear effusion isolates. Of these, 10 isolates were available from both middle ear and nasopharynx of the same patient. A *modM* gene was present in all isolates investigated, with *modM2* being the most common allele (83% of isolates), followed by allele 3 (16% of isolates), with allele 1 present in only one additional isolate (1%) (Table 3). Interestingly, a statistically significant association was found between *modM3* and middle ear isolates, with 47% of middle ear isolates containing *modM3* compared to 11% of nasopharyngeal isolates, and 53% of middle ear isolates containing *modM2*

compared to 87.5% of nasopharyngeal isolates (Table 3; Fisher's exact p-value = 0.0194). Furthermore, 2 of the 7 *modM3* nasopharyngeal isolates were obtained from patients who also had a *modM3* isolate from their middle ear. The complete list of isolates, their site of isolation and their *modM* allele, is shown in Table S1.

To investigate whether the nasopharyngeal and middle ear isolates from the same child were the same strain of *M. catarrhalis*, the total protein profiles from whole-cell lysates were compared by SDS-PAGE. Size differences between an abundant high molecular weight protein were particularly evident, suggesting that the over-representation of *modM3* strains in middle ear isolates was not due to a single predominating virulent strain. This major protein was identified as UspA2 by mass spectrometry. For 7 of the 9 pairs, the isolates had the same *modM* allele, suggesting the same strain was isolated (Fig. 5). For patients 150 and 159, however, strains containing different *modM* alleles were isolated from the different anatomic sites. It is notable that in each of these cases, a *modM3* isolate was in the middle ear rather than in the nasopharyngeal site. Additionally, patients 101 and 207 had the same *modM* allele from their respective middle ear and nasopharyngeal sites. However, protein profiles of nasopharyngeal and middle ear isolates indicated that for these 2 patients, despite the fact that both isolates had the same *modM* allele, they were nonetheless distinct strains of *M. catarrhalis* (Fig. 5).

Discussion

M. catarrhalis is a significant cause of otitis media in children, and exacerbations of chronic obstructive pulmonary disease in adults. These diseases have significant morbidity, with OM being the most common childhood bacterial infection and the main cause of pediatric antibiotic usage and visits to healthcare providers in the developed world (9, 10), and COPD the 5th leading cause of death worldwide (13). Currently, there is no *M. catarrhalis* vaccine available (32). Antibiotic resistance is common in *M. catarrhalis* (33) and antibiotic treatment is not able to deal with reinfection and the recurrent nature of *M. catarrhalis* associated disease. Understanding the biology of nasopharyngeal colonization, and the molecular mechanisms that underlie *M. catarrhalis* induced disease may enable the development of novel methods of treatment and prevention. Processes leading from asymptomatic colonization to overt disease are multifactorial, and may be both host- and microbe-dependent (34). Additionally, viral or bacterial co-infection is believed to play a role (reviewed in 35). For *M. catarrhalis*, however, there is no clear correlation between strains that cause specific disease and presence or expression of a particular virulence factor. Several adhesins and virulence factors have been identified in *M. catarrhalis* (reviewed in 2), including the phase variable outer membrane proteins UspA1 (36), UspA2H (37) and Hag/MID (38). Phase variation is

known to mediate bacterial adaptability and virulence (15). Phase variable DNA methyltransferases, which control the coordinated switching of expression of multiple genes (*i.e.,* a phase variable regulon, or phasevarion) (16), further enhance this adaptability. Phasevarions are increasingly being recognized as playing an important role in host-adapted pathogens. Phasevarions in *H. influenzae* (ModA) (17), the pathogenic *Neisseria* (ModA, ModB, ModD) (19, 20) and *H. pylori* (ModC) (21) control expression of surface antigens and virulence factors, leading to altered phenotypes between Mod ON and OFF variants. For example, the gonococcal ModA13 ON/OFF variants have distinct phenotypes for biofilm formation, resistance to antimicrobials, and survival in primary human cervical epithelial cells (19), the meningococcal ModA11 and ModA12 ON/OFF variants have altered susceptibility to certain antibiotics (39) and the meningococcal ModD1 ON / OFF variants have altered resistance to oxidative stress (20). Here we have characterized the ModM phase variable DNA methyltransferase of *M. catarrhalis.*

We identified three *modM* alleles in *M. catarrhalis*. Each allele differs most extensively within the putative DNA recognition domain. The presence of different alleles suggests that multiple phasevarions exist within *M. catarrhalis*, with each *modM* allele regulating a different set of genes. The allele most commonly observed in isolates, *modM2*, was characterized in *M. catarrhalis* strain 25239. Phase variation of *modM2* via the 5'-(CAAC)n-3' tetranucleotide repeats was confirmed, with variations in the 5'-CAAC-3' tract correlating with variable ModM2 expression. From a sample containing a mixed population of repeat lengths (18-22 repeats), screening of single colonies enabled the enrichment of three *M. catarrhalis* samples each containing a majority of either 19, 20 or 21 repeat tracts. ModM2 is expressed (ON) when 21 repeats are present, but not 19 or 20 repeats (OFF) due to a translational frameshift and the presence of premature stop codons in these frames. Homopolymeric A and T tracts are also present in the predicted promoter region, which vary in length between strains, suggesting the potential for phase variation. Therefore, in addition to the translational phase variation afforded by the 5'-(CAAC)n-3', *M. catarrhalis* may have evolved two additional mechanisms to phase vary expression of ModM at the transcriptional level. Further analysis is required to determine intra-strain variability of these homopolymeric tracts and their potential to modulate *modM* transcript levels.

Proteomic analysis revealed that 34 proteins were differentially expressed in *M. catarrhalis* grown on BHI media when ModM is ON compared to OFF, indicating that *modM* phase variation is involved in epigenetic regulation of proteins and mediates the coordinated switching of a phasevarion. The ModM2 phasevarion includes proteins involved in colonization and infection, including infection of the chinchilla nasopharynx and attachment to airway epithelial cells. Specifically, the following genes in the ModM phasevarion are differentially regulated or directly involved in: [1] infection of the chinchilla nasopharynx compared to growth in broth (25239_1211

(*miaE*), 25239_1678 (hypothetical), 25239_0666 (*fbpA*), 25239_1617 (*lebB*), 25239_1419 (*nqrC),* 25239_1611 (*typA*)) (40); [2] attachment to pharyngeal cells or alveolar cells (25239_0579 (*rpmG*), 25239_0666 (*ahcY*)) (41); [3] oxidative stress responses after exposure to H_2O_2 (25239_0666 (*fbpA*), 25239_0311(hypothetical), 25239_1095 (*greA*) and 25239_1027 (*bfrA*)) (42); and [4] growth in low iron conditions (25239_0666 (*fbpA*), 252399_0219 (*fixC*)) (43). The hypothetical protein encoded by 25239_1678 had decreased expression in *M. catarrhalis* 25239 ModM2 ON compared to ModM2 OFF. Interestingly its homologue in strain ATCC 43617 (MCORF1550) has a significant impact on infection potential in the chinchilla model, with the mutant having decreased fitness in a competitive chinchilla infection (40). MCORF1550 also has increased expression during attachment to human bronchial epithelial (HBE) cells (44). The combined effect of altered gene expression via the ModM phasevarion is predicted to have considerable effects on *M. catarrhalis* colonization and virulence in the human host, with the fittest subpopulation of ModM ON or OFF variants surviving as the bacteria transition through different host microenvironments. The expression profile analysis of the ModM2 phasevarion presented was only performed under one culture condition. However, there are expected to be niche specific differences in phasevarion gene expression. A difference in gene expression between ModM ON and OFF can only be detected if the genes in question are being expressed (*e.g.,* activated or de-repressed by regulatory proteins under specific conditions), and growth under different physiologically relevant conditions may reveal further virulence-associated genes under the control of the ModM2 phasevarion.

Whole genome SMRT sequencing and methylome analysis demonstrated that ModM2 is a functional m6A (N6-methyladenosine) methyltransferase that methylates the second adenine in 100% of 5'-GAR^{m6}AC-3' sequences in the genome. SMRT methylome analysis is a rapid method to identify methylation sites in genomes which is increasingly being used to study bacterial methylomes (28, 45, 46). ModM2 inhibition of digestion of genomic DNA was tested, and the expected inhibition of digestion with BsmBI was seen for ModM2 ON but not the ModM2 OFF and mutant strains, validating the SMRT identification of 5'-GAR^{m6}AC-3' as the substrate for ModM2 methylation. The precise mechanisms by which ModM2 methylation of DNA regulates differential gene transcription remain unclear but the phenomenon of epigenetic regulation in bacteria is well described, and DNA methylation can alter interaction of regulatory proteins with DNA-binding sites, which directly regulates transcription (24). The modified 5'-GARAC-3' site was identified upstream of several genes controlled by the ModM2 phasevarion (Fig. S2), and this information will facilitate investigations of the molecular mechanism of regulation by ModM2. However, epigenetic regulation may also be indirect and multifactorial, with expression of a regulator or a protein in a tightly regulated system being altered (*e.g.,* the ferric uptake regulator Fur or a protein

involved in iron homeostasis, respectively) resulting in altered expression of another protein in the same regulon. The level of differential expression seen for specific genes is also expected to vary based on the location of the methylation site and the transcription factor being affected. Furthermore, the percent ModM2 ON vs. OFF in the sample (>90% for the samples in this study) would affect the level of expression as measured for the population as a whole.

In terms of distribution, all *M. catarrhalis* strains examined to date contain a *modM* gene. Three *modM* alleles were identified (*modM1-3*) based on sequence analysis of the putative DNA-recognition domain. The *modM2* allele was found in 80% of strains, while *modM3* was present in 19% of strains investigated. *modM1* was first identified in strain ATCC 23246 (25), but was only found in one additional strain in this study. Furthermore, ATCC 23246 has been described as atypical and as being more closely related to *Moraxella canis* than *M. catarrhalis* (47).

Interestingly, a difference was found between the site of isolation (middle ear or nasopharynx) and the *modM* allele present. There is a statistically significant association between *modM3* and middle ear isolates. Despite only representing 19% of all isolates, *modM3* was over represented in middle ear isolates (47% of middle ear isolates). Conversely, only 11% of nasopharyngeal isolates were *modM3* strains (and of these 7 isolates, 2 were from the nasopharynx of children where a *modM3* was also isolated from the middle ear). We have subsequently determined that the ModM3 methylation site is 5'-AC^{m6}ATC-3', which is distinct to the methylation site of ModM2 suggesting that the ModM3 phasevarion comprises a different set of genes. Phasevarions have the potential to impact virulence via epigenetic regulation, and ModM3 is the only *M. catarrhalis* factor identified to date that is strongly associated with otitis media compared to nasopharyngeal carriage. We have reported a similar association in an independent study on *N. meningitidis* phasevarions, where the ModD1 DNA methyltransferase is almost exclusively found in the hypervirulent meningococcal clonal complex 41/44 (20). The proteins regulated and the phenotypes of ModM2 and ModM3 ON vs. OFF variants in different infection and virulence models are under investigation.

In summary, we have confirmed modulation of gene expression through the random ON/OFF switching of the ModM2 phasevarion, including several genes with demonstrated expression changes or effects in virulence models. This modulation of expression could aid *M. catarrhalis* adaptation to changing host environments, potentially contributing to increased fitness for transmission and/or ability to cause disease. The significant association between *modM3* and middle ear isolates requires further investigation, but suggests a role in pathogenesis.

Acknowledgements

This work was supported by the Australian National Health and Medical Research Council (NHMRC) [CDF to K.L.S., Program Grant 565526 to M.P.J.]. L.V.B is supported by an Australian Postgraduate Award. iTRAQ was performed at Australian Proteome Analysis Facility (APAF) the infrastructure provided by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

References

1. Vergison, A. (2008) Microbiology of otitis media: a moving target. *Vaccine* **26 Suppl 7**, G5-10
2. Murphy, T. F., and Parameswaran, G. I. (2009) Moraxella catarrhalis, a human respiratory tract pathogen. *Clin Infect Dis* **49**, 124-131
3. Post, J. C., Preston, R. A., Aul, J. J., Larkins-Pettigrew, M., Rydquist-White, J., Anderson, K. W., Wadowsky, R. M., Reagan, D. R., Walker, E. S., Kingsley, L. A., Magit, A. E., and Ehrlich, G. D. (1995) Molecular analysis of bacterial pathogens in otitis media with effusion. *JAMA* **273**, 1598-1604
4. Mackenzie, G. A., Leach, A. J., Carapetis, J. R., Fisher, J., and Morris, P. S. (2010) Epidemiology of nasopharyngeal carriage of respiratory bacterial pathogens in children and adults: cross-sectional surveys in a population with high rates of pneumococcal disease. *BMC infectious diseases* **10**, 304
5. Faden, H., Duffy, L., Wasielewski, R., Wolf, J., Krystofik, D., and Tung, Y. (1997) Relationship between nasopharyngeal colonization and the development of otitis media in children. Tonawanda/Williamsville Pediatrics. *The Journal of infectious diseases* **175**, 1440-1445
6. Revai, K., Mamidi, D., and Chonmaitree, T. (2008) Association of nasopharyngeal bacterial colonization during upper respiratory tract infection and the development of acute otitis media. *Clin Infect Dis* **46**, e34-37
7. Faden, H., Harabuchi, Y., and Hong, J. J. (1994) Epidemiology of Moraxella catarrhalis in children during the first 2 years of life: relationship to otitis media. *The Journal of infectious diseases* **169**, 1312-1317
8. A.A.P (2004) Diagnosis and management of acute otitis media. (AAP, American Academy of Pediatrics Subcommittee on Management of Acute OM). *Pediatrics* **113**, 1451-1465
9. Freid, V. M., Makuc, D. M., and Rooks, R. N. (1998) Ambulatory health care visits by children: principal diagnosis and place of visit. *Vital and health statistics. Series 13, Data from the National Health Survey*, 1-23
10. Gonzales, R., Malone, D. C., Maselli, J. H., and Sande, M. A. (2001) Excessive antibiotic use for acute respiratory infections in the United States. *Clin Infect Dis* **33**, 757-762
11. Sethi, S., and Murphy, T. F. (2008) Infection in the pathogenesis and course of chronic obstructive pulmonary disease. *N Engl J Med* **359**, 2355-2365
12. Murphy, T. F., Brauer, A. L., Grant, B. J., and Sethi, S. (2005) Moraxella catarrhalis in chronic obstructive pulmonary disease: burden of disease and immune response. *American journal of respiratory and critical care medicine* **172**, 195-199

13. Mannino, D. M., and Buist, A. S. (2007) Global burden of COPD: risk factors, prevalence, and future trends. *Lancet* **370**, 765-773
14. Revai, K., McCormick, D. P., Patel, J., Grady, J. J., Saeed, K., and Chonmaitree, T. (2006) Effect of pneumococcal conjugate vaccine on nasopharyngeal bacterial colonization during acute otitis media. *Pediatrics* **117**, 1823-1829
15. Moxon, R., Bayliss, C., and Hood, D. (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**, 307-333
16. 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. *Nature reviews. Microbiology* **8**, 196-206
17. 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. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5547-5551
18. 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
19. 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., Maiden, M. C., Grimmond, S. M., Apicella, M. A., and Jennings, M. P. (2009) Phasevarions mediate random switching of gene expression in pathogenic Neisseria. *PLoS Pathog* **5**, e1000400
20. Seib, K. L., Pigozzi, E., Muzzi, A., Gawthorne, J. A., Delany, I., Jennings, M. P., and Rappuoli, R. (2011) A novel epigenetic regulator associated with the hypervirulent Neisseria meningitidis clonal complex 41/44. *FASEB J* **25**, 3622–3633
21. 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
22. Gawthorne, J. A., Beatson, S. A., Srikhanta, Y. N., Fox, K. L., and Jennings, M. P. (2012) Origin of the diversity in DNA recognition domains in phasevarion associated modA genes of pathogenic Neisseria and Haemophilus influenzae. *PLoS One* **7**, e32337
23. Seib, K. L., Jen, F. E., Tan, A., Scott, A. L., Kumar, R., Power, P. M., Chen, L., Wu, H., Wang, A. H., Boitano, M., Clark, T. A., Korlach, J., Rao, D. N., and Jennings, M. P. (2014) Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N6-adenine DNA methyltransferases of Neisseria meningitidis. *Nucleic Acids Res* **submitted**

503 24. Wion, D., and Casadesus, J. (2006) N6-methyl-adenine: an epigenetic signal for DNA-
504 protein interactions. *Nature reviews. Microbiology* **4**, 183-192
505 25. Seib, K. L., Peak, I. R., and Jennings, M. P. (2002) Phase variable restriction-modification
506 systems in Moraxella catarrhalis. *FEMS Immunol Med Microbiol* **32**, 159-165
507 26. Wu, H. J., Seib, K. L., Srikhanta, Y. N., Edwards, J., Kidd, S. P., Maguire, T. L., Hamilton,
508 A., Pan, K. T., Hsiao, H. H., Yao, C. W., Grimmond, S. M., Apicella, M. A., McEwan, A.
509 G., Wang, A. H., and Jennings, M. P. (2010) Manganese regulation of virulence factors and
510 oxidative stress resistance in Neisseria gonorrhoeae. *J Proteomics* **73**, 899-916
511 27. Seib, K. L., Oriente, F., Adu-Bobie, J., Montanari, P., Ferlicca, F., Giuliani, M. M.,
512 Rappuoli, R., Pizza, M., and Delany, I. (2010) Influence of serogroup B meningococcal
513 vaccine antigens on growth and survival of the meningococcus in vitro and in ex vivo and in
514 vivo models of infection. *Vaccine* **28**, 2416-2427
515 28. Clark, T. A., Murray, I. A., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M.,
516 Fomenkov, A., Roberts, R. J., and Korlach, J. (2012) Characterization of DNA
517 methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic*
518 *Acids Res* **40**, e29
519 29. Song, X., Bandow, J., Sherman, J., Baker, J. D., Brown, P. W., McDowell, M. T., and
520 Molloy, M. P. (2008) iTRAQ experimental design for plasma biomarker discovery. *Journal*
521 *of proteome research* **7**, 2952-2958
522 30. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion
523 for mass spectrometric characterization of proteins and proteomes. *Nature protocols* **1**,
524 2856-2860
525 31. Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2010) REBASE--a database for DNA
526 restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **38**, D234-236
527 32. Murphy, T. F. (2009) Vaccine development for Moraxella catarrhalis: rationale, approaches
528 and challenges. *Expert Rev Vaccines* **8**, 655-658
529 33. Bootsma, H. J., van Dijk, H., Vauterin, P., Verhoef, J., and Mooi, F. R. (2000) Genesis of
530 BRO beta-lactamase-producing Moraxella catarrhalis: evidence for transformation-mediated
531 horizontal transfer. *Molecular microbiology* **36**, 93-104
532 34. Murphy, T. F. (1996) Branhamella catarrhalis: epidemiology, surface antigenic structure,
533 and immune response. *Microbiological reviews* **60**, 267-279
534 35. Bakaletz, L. O. (2010) Immunopathogenesis of polymicrobial otitis media. *J Leukoc Biol*
535 **87**, 213-222

536 36. Lafontaine, E. R., Wagner, N. J., and Hansen, E. J. (2001) Expression of the Moraxella
537 catarrhalis UspA1 protein undergoes phase variation and is regulated at the transcriptional
538 level. *Journal of bacteriology* **183**, 1540-1551
539 37. Wang, W., Pearson, M. M., Attia, A. S., Blick, R. J., and Hansen, E. J. (2007) A UspA2H-
540 negative variant of Moraxella catarrhalis strain O46E has a deletion in a homopolymeric
541 nucleotide repeat common to uspA2H genes. *Infection and immunity* **75**, 2035-2045
542 38. Mollenkvist, A., Nordstrom, T., Hallden, C., Christensen, J. J., Forsgren, A., and Riesbeck,
543 K. (2003) The Moraxella catarrhalis immunoglobulin D-binding protein MID has conserved
544 sequences and is regulated by a mechanism corresponding to phase variation. *Journal of*
545 *bacteriology* **185**, 2285-2295
546 39. Jen, F. E., Seib, K. L., and Jennings, M. P. (2014) Phasevarions mediate epigenetic
547 regulation of antimicrobial susceptibility in Neisseria meningitidis. *Antimicrob Agents*
548 *Chemother* **In press.**
549 40. Hoopman, T. C., Liu, W., Joslin, S. N., Pybus, C., Sedillo, J. L., Labandeira-Rey, M.,
550 Laurence, C. A., Wang, W., Richardson, J. A., Bakaletz, L. O., and Hansen, E. J. (2012) Use
551 of the chinchilla model for nasopharyngeal colonization to study gene expression by
552 Moraxella catarrhalis. *Infection and immunity* **80**, 982-995
553 41. de Vries, S. P., Eleveld, M. J., Hermans, P. W., and Bootsma, H. J. (2013) Characterization
554 of the molecular interplay between Moraxella catarrhalis and human respiratory tract
555 epithelial cells. *PLoS One* **8**, e72193
556 42. Hoopman, T. C., Liu, W., Joslin, S. N., Pybus, C., Brautigam, C. A., and Hansen, E. J.
557 (2011) Identification of gene products involved in the oxidative stress response of Moraxella
558 catarrhalis. *Infection and immunity* **79**, 745-755
559 43. Wang, W., Reitzer, L., Rasko, D. A., Pearson, M. M., Blick, R. J., Laurence, C., and
560 Hansen, E. J. (2007) Metabolic analysis of Moraxella catarrhalis and the effect of selected in
561 vitro growth conditions on global gene expression. *Infection and immunity* **75**, 4959-4971
562 44. Wang, W., Joslin, S. N., Pybus, C., Evans, A. S., Lichaa, F., Brautigam, C. A., and Hansen,
563 E. J. (2014) Identification of an Outer Membrane Lipoprotein Involved in Nasopharyngeal
564 Colonization by Moraxella catarrhalis in an Animal Model. *Infection and immunity*
565 45. Murray, I. A., Clark, T. A., Morgan, R. D., Boitano, M., Anton, B. P., Luong, K.,
566 Fomenkov, A., Turner, S. W., Korlach, J., and Roberts, R. J. (2012) The methylomes of six
567 bacteria. *Nucleic Acids Res* **40**, 11450-11462
568 46. Krebes, J., Morgan, R. D., Bunk, B., Sproer, C., Luong, K., Parusel, R., Anton, B. P., Konig,
569 C., Josenhans, C., Overmann, J., Roberts, R. J., Korlach, J., and Suerbaum, S. (2014) The

complex methylome of the human gastric pathogen Helicobacter pylori. *Nucleic Acids Res* **42**, 2415-2432

47. Juni, E., and Bovre, K. (2007) Family II. Moraxellaceae. In *Bergey's Manual of Systematic Bacteriology, Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria* (Garrity, G., Brenner, D. J., Krieg, N. R., and Staley, J. R., eds), Springer

48. Davie, J. J., Earl, J., de Vries, S. P., Ahmed, A., Hu, F. Z., Bootsma, H. J., Stol, K., Hermans, P. W., Wadowsky, R. M., Ehrlich, G. D., Hays, J. P., and Campagnari, A. A. (2011) Comparative analysis and supragenome modeling of twelve Moraxella catarrhalis clinical isolates. *BMC Genomics* **12**, 70

49. de Vries, S. P., van Hijum, S. A., Schueler, W., Riesbeck, K., Hays, J. P., Hermans, P. W., and Bootsma, H. J. (2010) Genome analysis of Moraxella catarrhalis strain BBH18, [corrected] a human respiratory tract pathogen. *Journal of bacteriology* **192**, 3574-3583

50. Zomer, A., de Vries, S. P., Riesbeck, K., Meinke, A. L., Hermans, P. W., and Bootsma, H. J. (2012) Genome sequence of Moraxella catarrhalis RH4, an isolate of seroresistant lineage. *Journal of bacteriology* **194**, 6969

586 **TABLES AND FIGURES**

587 **Table 1:** Distribution of *modM* alleles in *M. catarrhalis* genome strains

Strain	**Clinical details of original isolate**	**Accession number [coordinates][a] (ref)**	**CAAC repeats[b]**	***modM* allele[c]**
ATCC 23246	NA	AY049056 [1755-3665] (25)	20	1
ATCC 25239	NA	CP007669 (this study)	21	2
BC8	Sinus Wash, Sinusitis; Child	AERJ0100020.1 [28061-30020] (48)	22	2
CO72	Middle ear fluid (OME); Child	AERK01000020.1/ AERK0100005.1 [27689-87834/ 1-1971] (48)	ND	2
O35E	OME; Child	AERL0100013.1 [27809-29743] (48)	30	2
103P14B1	COPD Exacerbation; Adult	AERE01000085.1/AERE01000006.1 [27451-27518/ 273821-275740] (48)	ND	2
BC7	Middle Ear, OME; Child	AERI01000013.1 [306295-308218] (48)	13	2
12P80B1	COPD Exacerbation; Adult	AERG01000018.1/AERG01000016.1 [1-46 /73283-75176] (48)	ND	2
46P47B1	COPD Exacerbation; Adult	AERF0100013.1/ AERF0100018.1 [27564-27619/ 219077-220982] (48)	ND	2
7169	OME; Child	AERC01000018.1/ AERC01000019.1 [53144-53185/ 1-1878] (48)	ND	2
BBH18	sputum of a patient with COPD during an exacerbation	CP002005.1 [375514-377490] (49)	30	2
ATCC 43617	Transtracheal aspirate, chronic bronchitis; Coal miner	AX067463 [54028-52979] (World patent WO0078968)	19	2
RH4	Blood	NC_014147.1 [375514-377490] (50)	30	2
BC1	Tracheal Aspirate, Bronchiolitis; Child	AERH0100007.1 [27691-29656] (48)	31	3
101P30B1	COPD Exacerbation;	AEPC01000023.1/ AEPC01000009.1	ND	3

	Adult	[53095-53191/ 305329-307249] (48)		

588 [a] Genome accession number is shown followed by the coordinates of the *modM* gene.

589 [b] Number of 5'-(CAAC)-3' repeats within the *modM* gene. ND, repeat number not determined due

590 to a gap between contigs in the genome sequence.

591 [c]*modM* alleles determined based on ≥95% identity of the DNA recognition domain (see Fig.

592 1*B/C* for strains and sequences that define the *modM* alleles).

593 NA, not available.

594

595 **Table 2.** Differentially expressed proteins from iTRAQ studies of *M. catarrhalis* 25239 ModM2
596 ON and OFF strains

Gene symbol [a]	Name	Role category (location)[b]	Fold change ON:OFF[c]
Increased expression in ModM2 ON			
0380	DNA methyltransferase	1 (C)	4.4
1724	Acyl-CoA thioesterase 2	2 (C)	1.5
0454	Penicillin-binding protein 1A, *pbp1A*	3 (IM/P)	1.4
1430	Hypothetical protein	4 (C)	1.4
1390	Hypothetical protein	5 (IM/P)	1.4
1486	50S ribosomal protein L32, *rpmF*	6 (C)	1.4
1201	Cytochrome c biogenesis protein CcmG, *dsbE*	4 (IM/P)	1.4
0563	LemA family protein	7 (OM)	1.4
0316	putative binding protein component of ABC iron transporter, *fbpA*	8 (IM/P)	1.4
0382	L-2,4-diaminobutyrate decarboxylase	9 (C)	1.4
0219	Electron transfer flavoprotein-ubiquinone oxidoreductase, *fixC*	10 (C)	1.4
1617	Signal peptidase I, *lepB*	4 (IM)	1.3
1419	Na(+)-translocating NADH-quinone reductase subunit C, *nqrc*	8 (OM)	1.3
1531	30S ribosomal protein S21, *rspu*	6 (C)	1.3
0039	Oxa1Ec, *yidC*	4 (IM)	1.3
0128	Chaperone protein, *dnaJ*	4 (C)	1.3
0666	Adenosylhomocysteinase, *ahcY*	10 (C)	1.3
1493	30S ribosomal protein S11, *rpsK*	6 (C)	1.3
0579	50S ribosomal protein L33, *rpmG*	6 (C)	1.3
Decreased expression in ModM2 ON			**OFF:ON**
0826	Putative chromosome-partitioning protein, *parB*	7 (C)	1.3
1451	Hypothetical protein	5 (EC)	1.3
1244	Glycine cleavage system H protein, *gcvH*	10 (C)	1.3
1027	Bacterioferritin, *bfrA*	8 (C)	1.3
0374	Guanylate kinase, *gmk*	11 (C)	1.3
0311	Hypothetical protein	5 (C)	1.3
1611	Tyrosine phosphorylated protein A, *typA*	12 (C)	1.3
1095	Transcript cleavage factor, *greA*	13 (C)	1.4
1677	Hypothetical protein, *rbfA*	13 (C)	1.4
0995	Hypothetical protein	5 (IM)	1.4
1312	Glucose-inhibited division protein A, *gidA*	6 (C)	1.4
181	23S rRNA (uracil(1939)-C(5))-methyltransferase RlmD, *rumA*	6 (C)	1.5
1211	tRNA-(MS[2]IO[6]A)-hydroxylase, *miaE*	6 (C)	1.5
1678	Hypothetical protein	5 (C)	1.7
1483	Acyl carrier protein	5 (C)	1.9

597 [a] Gene symbols are gene annotations from *M. catarrhalis* strain 25239 (GenBank accession number
598 CP007669).

599 [b] Assigned role category and predicted protein location from de Vries *et al.,* (49). Role categories:
600 (1) DNA metabolism; (2) Fatty acid and phospholipid metabolism; (3) Cell envelope; (4) Protein
601 fate; (5) Hypothetical proteins; (6) Protein synthesis; (7) Unknown function; (8) Transport and
602 binding proteins; (9) Central intermediary metabolism; (10) Energy metabolism; (11) Purines.
603 pyrimidines. nucleosides and nucleotides; (12) Cellular processes; (13) Transcription.
604 Predicted location: (C) cytoplasm; (IM) inner membrane; (IM/P) inner membrane/periplasm; (OM)
605 outer membrane; (EC) extracellular.
606 [c] Expression ratio of ≥ 1.3-fold between the ModM2-ON/OFF variants, with a P-value of ≤0.05.
607

608 **Table 3:** Distribution of *modM* alleles in a panel of 81 *M. catarrhalis* isolates from the
609 nasopharynx or middle ear of children.

Allele	**Overall % (n)**	**Nasopharyngeal % (n)**	**Middle Ear % (n)**
modM1	1 (1)	1 (1)	0 (0)
modM2	80 (65)	88 (56)	53 (9)
modM3	19 (15)	11 (7)	47 (8)*
Total	100 (81)	100 (64)	100 (17)

610 * Association of *modM3* with middle ear isolates. Fisher exact p value= 0.0194

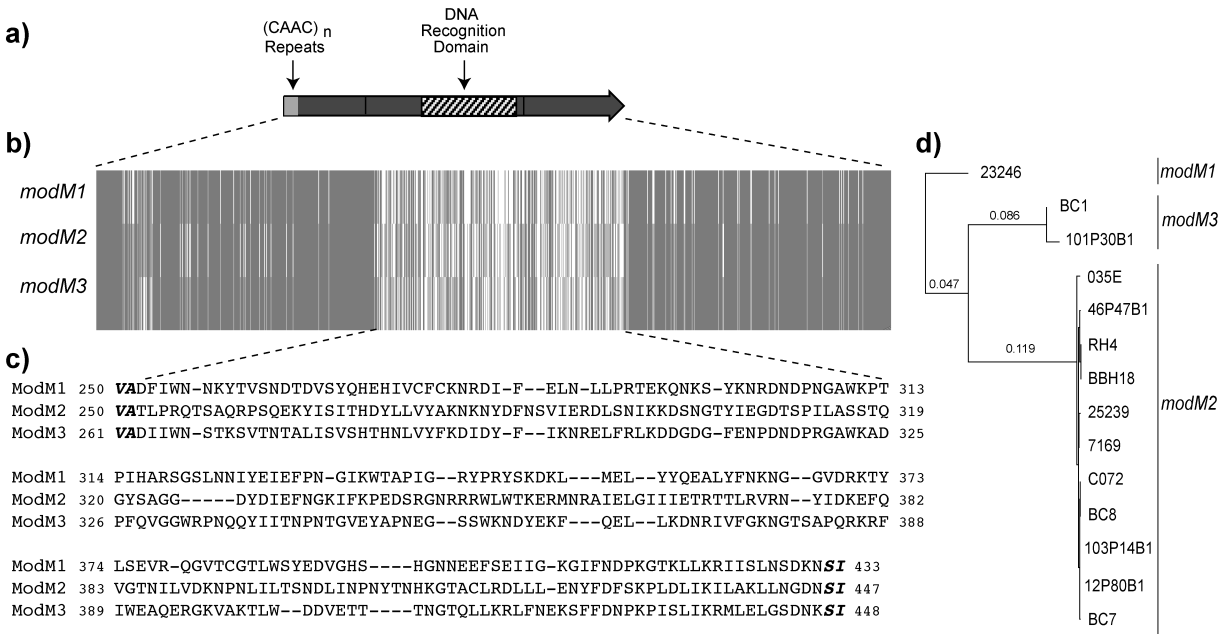

Figure 1. ***M. catarrhalis modM*** **alleles. (A)** Schematic representation of the methyltransferase (*modM*) gene showing the 5'-$(CAAC)_n$-3' DNA repeat tract responsible for phase variable expression, and the variable central DNA recognition domain (DRD). **(B)** *modM1–3* alleles were aligned in ClustalW and visualized with JalView [identical nucleotides are shown as vertical lines, with identity over a run of nucleotides shown as dark gray (>80% identity), light gray (>50%), or white (<50% identity or a gap)]. *M. catarrhalis* strains that define the *modM* alleles are ATCC 23246 (M1), ATCC 25239 (M2) and BC1 (M3), with a strain's *modM* allele is classified based on ≥95% nucleotide identity of the DNA recognition domain to one of these sequences (see Table 1). **(C)** The predicted amino acid sequences of the DRD of ModM1-3 are shown, as well as two conserved amino acids shown in bold on each side of the variable region. **(D)** Homology between the *modM* alleles is shown, using a phylogenetic reconstruction based on a neighbor-joining distance matrix.

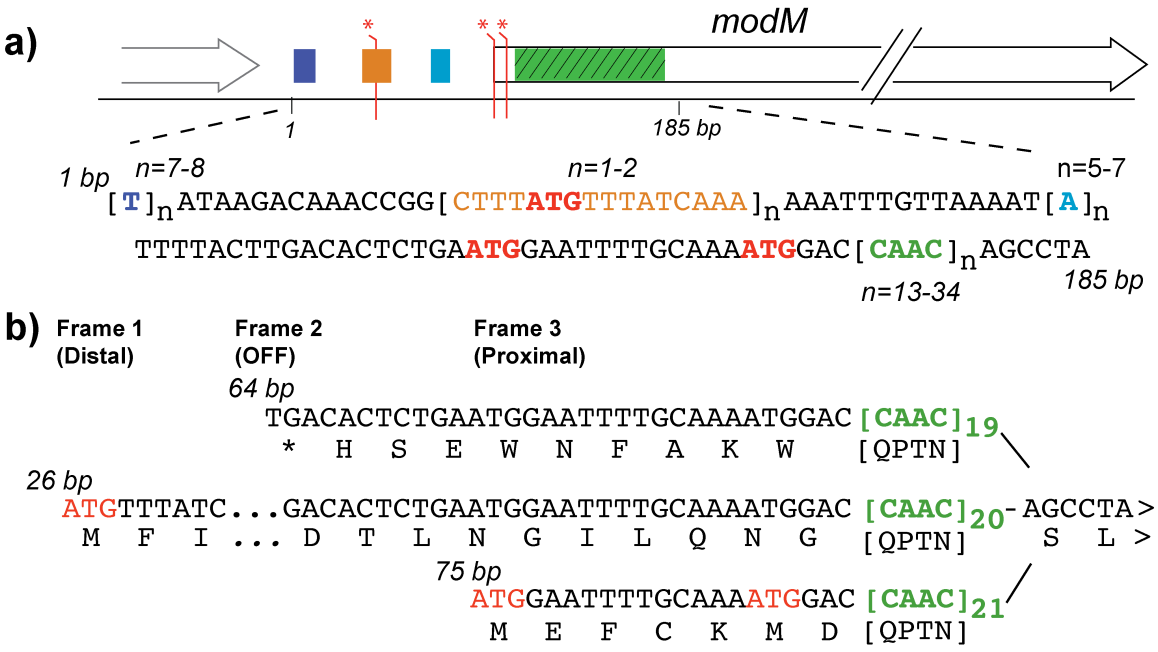

Figure 2. DNA repeats present in the *modM* locus. (A) Schematic of the *modM* methyltransferase showing the location of the 3 potential ATG translational start sites (red/asterisk) and the DNA repeats in the intergenic region and open reading frame; poly T tract (purple), direct repeats (orange), poly A tract (light blue), and tetranucleotide 5'-$(CAAC)_n$-3' repeats (green/diagonal lines). Below, the nucleotide sequence of this region is shown. The range of repeats seen between strains is shown above each repeat unit. The numbering, 1-185 bp, refers to *M. catarrhalis* strain 25239, starting from the poly T tract. **(B)** Schematic of the *modM2* locus from strain 25239 showing that translation of the *mod* gene could be initiated from two of three frames (Distal or Proximal, but not OFF) depending on the number of 5'-CAAC-3' repeats. Asterisk indicates stop codon. Numbering is as described above.

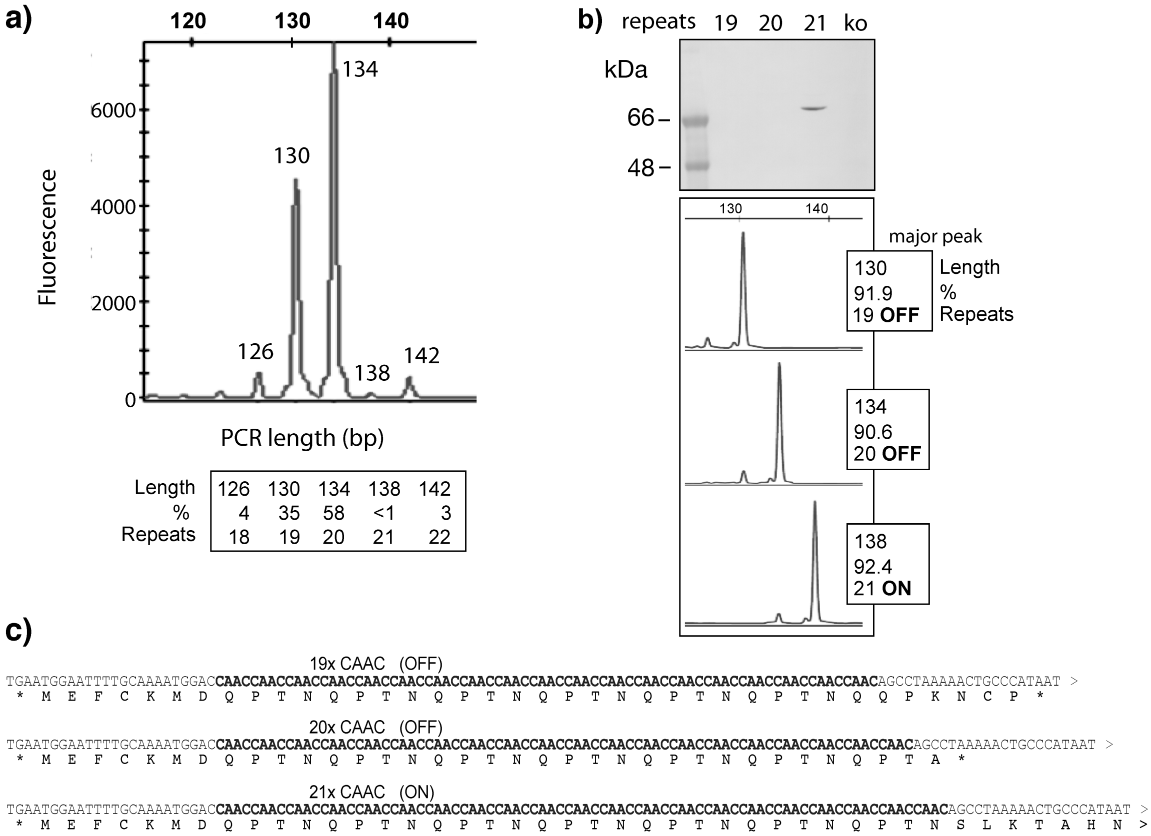

Figure 3. Phase variable expression of *modM2*. (A) Fluorescent peaks observed in GeneScan fragment length analysis of a heterogeneous *M. catarrhalis* sample containing 18-22 repeats, with the fragment length, its relative percentages within the sample, and corresponding number of 5'-(CAAC)$_n$-3' tetranucleotide repeats indicated below. **(B)** Western blot analysis confirming ModM2 expression in whole cell lysates of samples, isolated from screening single colonies, containing 21 CAAC repeats (but not 19 or 20) and the absence of ModM2 expression in the *modM2::kan* knockout (ko) strain. Results of GeneScan fragment length analysis of the DNA repeat region in the samples used in the Western blot is shown below, confirming >90% of each population contained either 19, 20 or 21 repeats. **(C)** Putative ModM translation is shown for 19 (OFF), 20 (OFF), and 21 (ON) repeat units from the proximal initiation codons. Asterisk indicates stop codon.

a)

Methyltransferase Specificity	Modified Base	*modM2* ON	*modM2* OFF	*modM2* KO	Assignment[#]
5'-GAR**A**C-3'	m6A	100%	0%	0%	M.McaM2 (ModM2)
5'-AC**A**N_6CTTG-3' 3'-TGTN_6G**A**AC-5'	m6A	100% 100%	100% 99.7%	94.5% 90.7%	ND
5'-GCNG**C**-3' 3'-**C**GNCG-5'	m5C	*	*	*	M.Mca 25239_00017

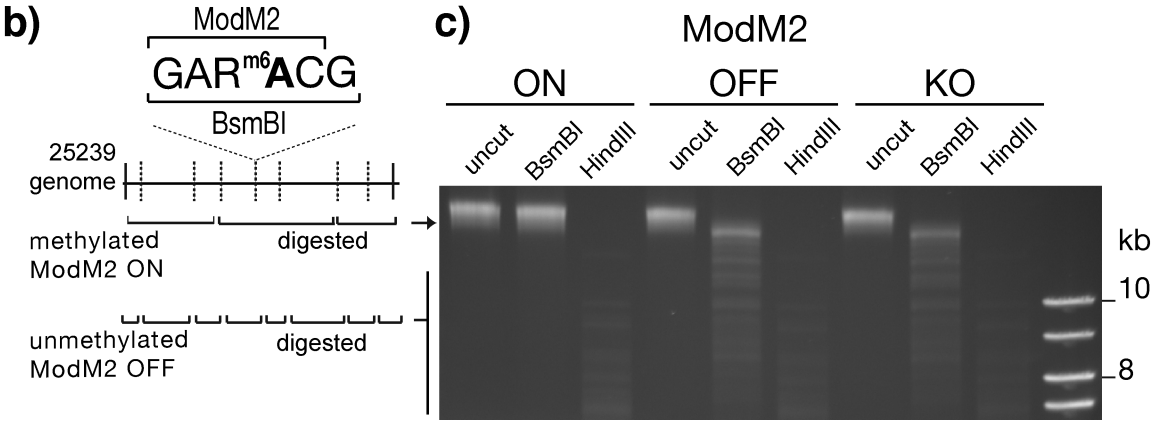

Figure 4: Methylome analysis of *M. catarrhalis* 25239 and ModM2 methylation of 5'-GAR^{m6}AC-3'. (A) SMRT methylome analysis of *M. catarrhalis* 25239 *modM2* ON, OFF and *modM2::kan* mutant strains. The % of recognition sites that are detected as methylated in the genome is shown for m6A methylation. * Due to the weak SMRT kinetic signal of 5mC methylation, % of sites methylated has not been calculated. [#]Assignment is based on direct evidence (M.McaM2 (ModM2)) or homology (M.Mca25239_00017). ND, not determined; no methyltransferase could be reliably assigned as responsible for methylation of this site. **(B)** Schematic of the restriction-inhibition assay. An example is shown of one of the many ModM2 methylation sites (5'-GAR^{m6}AC-3') that overlaps the methyl-sensitive BsmBI restriction enzyme site (5'-GARACG-3') in the genome of *M. catarrhalis* 25239, and below are diagrammatical representations of the methylated and unmethylated genomes. BsmBI does not cut if the adenosine is methylated in the overlapping ModM2 methylation site ('methylated' represents ModM2 ON digested DNA), but will digest DNA that is not methylated at this site ('unmethylated' represents ModM2 OFF digested DNA). Approximately 35% of BsmBI sites overlap with the ModM2 site in the 25239 genome. **(C)** DNA agarose gel of uncut, BsmBI and HindIII digested genomic DNA from ModM2 ON, ModM2 OFF and Δ*modM2::kan* (ko) strains, where methylated DNA in the ModM2 ON strain is protected from BsmBI digestion. HindIII was used as an unrelated positive control for DNA digestion.

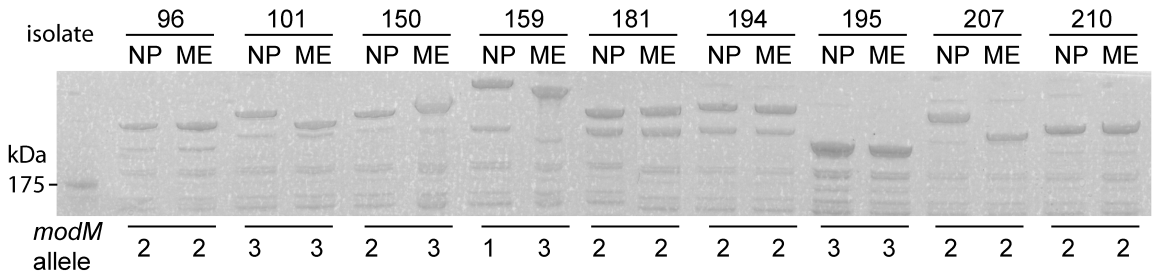

Figure 5: SDS-PAGE of whole cell lysates of paired *M. catarrhalis* isolates.

The isolate number and site of isolation (NP, nasopharynx; ME, middle ear) of paired isolate (*i.e.,* isolates from the same child) is shown above the coomassie stained polyacrylamide gel, and the *modM* allele present is shown below the gel. The predominant high molecular weight protein was identified by mass spectrometry as UspA2.

Supplementary data.

Table S1. ModM alleles in *M. catarrhalis* strains investigated

Strain number	Site of isolation *	*modM* allele
8	NPL	2
9	NPR	2
12	NPL	2
18	NPL	2
24	NPL	2
27	NPR	2
30	NPR	2
31	NPR	2
37	NPR	2
45	NPR	2
46	NPL	2
50	***MER***	***2***
55	NPR	3
56	NPR	2
60	NPR	2
70	NPR	2
73	NPL	2
74	NPL	2
86	NPL	2
89	NPL	2
92	NPR	2
96	**NPL**	**2**
96	***MER***	***2***
97	NPR	2
99	NPL	2
101	**NPR**	**3**
101	***MER***	***3***
103	NPR	3
117	NPL	2
122	NPL	2
123	NPR	2
133	NPR	2
141	NPR	2
143	***MER***	***3***
147	***MER***	***2***
148	NPL	2
150	**NPR**	**2**
150	***MER***	***3***
153	NPL	2
155	NPR	2
156	NPR	2
157	NPR	2
159	**NPR**	**1**
159	***MEL***	***3***
166	NPR	2
171	NPL	3
172	NPL	2
173	NPR	2
174	NPL	2
176	NPR	2
181	***MER***	***2***
181	**NPL**	**2**
185	NPR	2
194	**NPL**	**2**
194	***MER***	***2***
195	**NPR**	**3**
195	***MER***	***3***
199	NPL	2
205	NPR	2
207	**NPR**	**2**
207	***MER***	***2***
210	**NP**	**2**
210	***MER***	***2***
304	NPL	2
352	NPR	2
363	NPR	2
367	NPL	2
380	NPR	2
383	NPR	2
396	NPR	2
397	NPR	2
398	NPR	3
401	NPR	3
405	NPL	2
411	NPR	2
417	NPL	2
42	***ME***	***3***
1857	***ME***	***3***
1879	***ME***	***2***
1889	***ME***	***3***
87069	***MEL***	***2***

* NP, nasopharynx; ME, middle ear; L, left; R, right. ME isolates are shown in bold italics, and paired isolates are shown in bold.

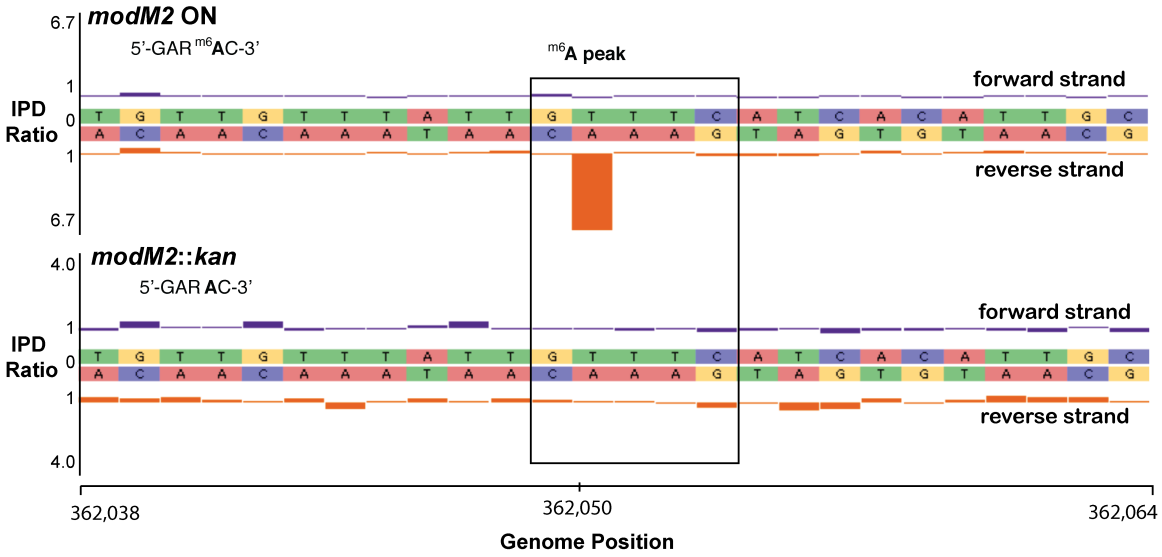

Figure S1. SMRT kinetic variation analysis and ModM2 methylation determination. Example trace of SMRT kinetic variation, showing the interpulse duration (IPD), which yields information regarding DNA methylation. The region of the *M. catarrhalis* 25239 genome shown is upstream of *fbpA*, which is differentially expressed between *modM2* ON and OFF. The 5'-GAR^{m6}AC-3' site (on the reverse strand) is methylated in *modM2* ON but not the OFF or *modM2::kan* mutant strain.

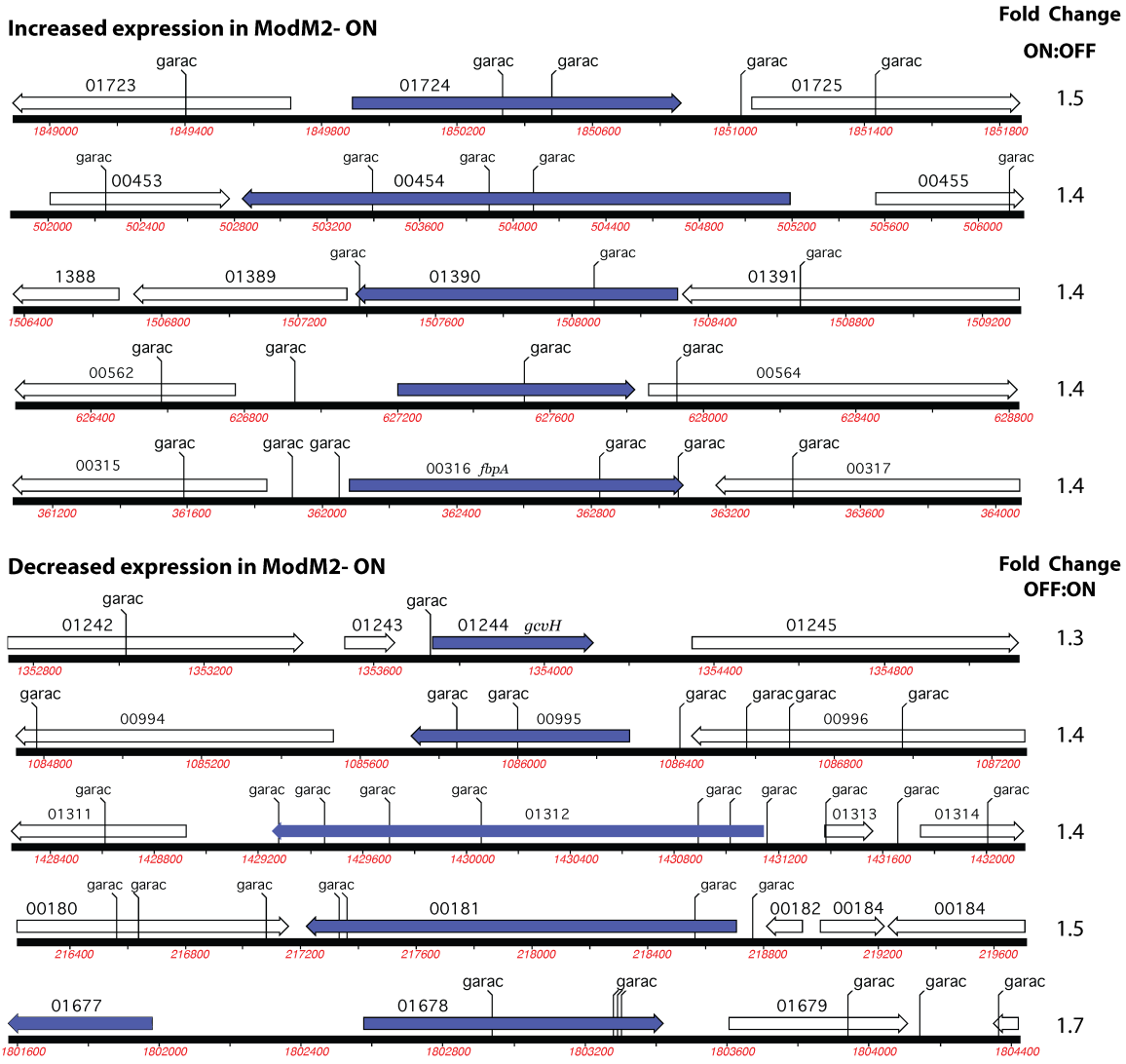

Figure S2. The position of ModM2 recognition sites with respect to a selection of genes encoding regulated proteins in the ModM2 phasevarion. Genes regulated in ModM2 ON vs. OFF strains are shaded. Methylated 5'-GAR^{m6}AC-3' sites are shown within the gene or within 1kb flanking regions of the regulated gene. The fold change in protein expression between ModM2 ON vs. OFF, as determined by iTRAQ, is shown on the right.