The Evolutionary History of Archaeal MCM Helicases: A Case Study of Vertical Evolution Combined with Hitchhiking of Mobile Genetic Elements

Mart Krupovič,1 Simonetta Gribaldo,2 Dennis H. Bamford,1 and Patrick Forterre*,2,3

1Department of Biosciences and Institute of Biotechnology, University of Helsinki, Helsinki, Finland
2Unité de Biologie Moléculaire du Genre chez les Extrémophiles, Département de Microbiologie, Institut Pasteur, Paris cedex, France
3Institut de Génétique et Microbiologie, Univ Paris-Sud, CNRS UMR8621, Orsay Cedex, France

*Corresponding author: E-mail: forterre@pasteur.fr.
Associate editor: Manolo Gouy

Abstract

Genes encoding DNA replication proteins have been frequently exchanged between cells and mobile elements, such as viruses or plasmids. This raises potential problems to reconstruct their history. Here, we combine phylogenetic and genomic context analyses to study the evolution of the replicative minichromosome maintenance (MCM) helicases in Archaea. Several archaeal genomes encode more than one copy of the mcm gene. Genome context analysis reveals that most of these additional copies are encoded within mobile elements. Exhaustive analysis of these elements reveals diverse groups of integrated archaeal plasmids or viruses, including several head-and-tail proviruses. Some MCMs encoded by mobile elements are structurally distinct from their cellular counterparts, with one case of novel domain organization. Both genome context and phylogenetic analysis indicate that MCM encoded by mobile elements were recruited from cellular genomes. An accelerated evolution and a dramatic expansion of methanococcal MCMs suggest a host-to-virus-to-host transfer loop, possibly triggered by the loss of the archaeal initiator protein Cdc6 in Methanococcales. Surprisingly, despite extensive transfer of mcm genes between viruses, plasmids, and cells, the topology of the MCM tree is strikingly congruent with the consensus archaeal phylogeny, indicating that mobile elements encoding mcm have coevolved with their hosts and that DNA replication proteins can be also useful to reconstruct the history of the archaeal domain.

Key words: MCM helicase, archaea, DNA replication, mobile elements.

Introduction

Cellular genomes are composed of genes from different origins (Koonin and Wolf 2008; Cortez et al. 2009). Some genes have been inherited vertically for a long period of time, whereas others have been more recently obtained in a particular cellular lineage following gene duplication, chromosome recombination (domain shuffling), or other processes linked to DNA replication. Finally, many genes have been transferred and fixed over time from different cellular or viral lineages. Integrated mobile elements (viruses, plasmids) appear indeed to be the major source of most recently transferred genes in cellular genomes (Cortez et al. 2009; Siefert 2009; Halary et al. 2010). Moreover, although most genes present in mobile elements are ORFans or have only homologues in closely related viruses and/or plasmids, mobile elements also often encode proteins that have been recruited by accident from cellular genomes and maintained if they had some advantage for virus and/or plasmid propagation. These genes can be transferred back to cellular lineages by homologous or illegitimate recombination between the mobile DNA and the chromosome of the host or else by integration of the entire mobile element into the cellular genome. Therefore, the presence of multiple homologues in a given genome, although usually considered as testifying for gene duplication or horizontal gene transfer (HGT) from other cells, could be also due to acquisition of an extra copy brought along by an integrated mobile element.

Genes encoding cellular DNA replication proteins are more likely than others to be recruited by mobile elements because they can be immediately useful for the replication of DNA viruses or plasmids. Hitchhiking of cellular genes by plasmids or viruses can accelerate their evolutionary rate. Later integration of such genes in cellular genomes will mimic gene duplication, whereas multiple integrations will produce gene expansion. It is also widely believe that mobile elements can easily cross barriers between distantly related evolutionary lineages, promoting extensive HGT of cellular genes. Gene duplication, fast evolution, and HGT can theoretically create insuperable problems for phylogenetic analyses. A possibility to bypass these pitfalls would be to identify genes that are present within mobile elements by analyzing their genomic environment to distinguish them from truly "cellular" genes. A gene introduced recently by a mobile element can be identified by the presence of characteristic genomic signatures such as the proximal integrase-coding genes, direct repeats overlapping with transfer RNA (tRNA) genes, homologues of other viral and/or plasmid genes, or else atypical nucleotide composition (Reiter et al. 1989; She et al. 2004; Cortez et al. 2009).
In this study, we have applied the above strategy to reconstruct the history of multiple MCM helicase genes in archaeal genomes. Indeed, recent observation that some archaeal viruses (Pagaling et al. 2007), proviruses (Krupovic and Bamford 2008a; Krupovic et al. 2010), and plasmids (Greve et al. 2005) encode replicative MCM helicasies suggests that some of these additional genes might have been also introduced in archaeal genomes by mobile elements (McGeoch and Bell 2008).

MCM proteins were first discovered in Saccharomyces cerevisiae as proteins essential for the maintenance of budding yeast minichromosomes (Maine et al. 1984). Later on, it was found that there are six distinct MCM paralogs, MCM2–7, that are conserved not only in yeast but also in all eukaryotic organisms. The six MCM proteins were found to form a heterohexameric complex acting as a replicative helicase by unwinding the duplex DNA to provide the single-stranded template for the DNA polymerase (Forsburg 2004). The archaeal initiation and elongation phases of chromosomal replication also rely on the action of a hexameric MCM helicase complex (Barry and Bell 2006). Strikingly, different archaeal species have a variable number of copies of MCM-encoding genes (Barry and Bell 2006; McGeoch and Bell 2008; Walters and Chong 2009). The number of MCM genes is especially high in Methanococcales and some of them clearly appear to be located within integrated mobile elements (Walters and Chong 2010). This suggested that some paralogous MCM in other archaeal orders could be also of viral/plasmidic origin and that studying the phylogeny and origin of archaeal MCM proteins could be a good starting point to evaluate the impact of the interaction between cells and viruses on the history of DNA replication proteins. Furthermore, considering the importance of MCM in both archaeal and eukaryotic DNA replication, knowledge of their history appears to be a prerequisite for their meaningful functional and biochemical analysis.

We have thus performed an in-depth phylogenetic and genomic context analysis of all archaeal mcm genes. Our results suggest that in most archaea with multiple mcm genes, only one of them is likely to produce a protein involved in replication of the cellular chromosome, whereas additional copies are encoded within potentially mobile elements. In particular, our data indicate that mobile elements have played a major role in the expansion of the MCM protein family in the order Methanococcales. Both genome context and phylogenetic analyses indicate that cellular mcm genes were recruited several times independently by unrelated mobile elements on multiple independent occasions. The complex phylogeny of multiple MCMs from Methanococcales suggests an evolutionary scenario including a host-to-virus-to-host transfer loop, possibly triggered by the loss of the archaeal initiator protein Cdc6 in Methanococcales. Some MCMs encoded by mobile elements are now structurally distinct from their cellular counterparts and might therefore deviate from them functionally. We identify in particular a group of MCM proteins from Methanococcales in which the C-terminal winged helix-turn-helix (wHTH) domain has been replaced by a similar but nonhomologous wHTH domain in the N-terminus. The MCM phylogeny does not show evidence for gene transfer between cellular lineages, indicating that mobile elements bearing mcm genes were not involved in gene transfer between archaeal lineages. In fact, the general topology of the MCM tree is strikingly congruent with the consensus archaeal tree. This indicates that cellular MCMs can be included in the set of core proteins useful to reconstruct archaeal phylogeny.

Materials and Methods

Collection of the Archaeal MCM Sequence Data Set
Archaeal MCM sequences were retrieved by running BlastP searches (BLOSUM62 matrix, 1 × 10−5 as an e-value cutoff; Altschul et al. 1997) using several randomly selected archaeal MCM sequences as seeds. The searches were limited to archaeal protein sequences available at the nonredundant protein database at National Center for Biotechnology Information (NCBI; 28 January 2009). The obtained data set was then analyzed to ensure that all archaeal species for which genome sequences are available were represented in the final MCM data set.

Criteria Used to Ascribe a Genomic Region as Being an Integrated Mobile Element or Provirus
Genomic sequences of archaeal species with more than one mcm gene copy per chromosome were analyzed for the presence of integrated genetic elements using CLC Main Workbench 5.0 software package (CLC Bio, Inc.). mcm gene–containing region was considered to be potentially mobile when two criteria were satisfied: 1) integrase-coding gene was present in proximity to the mcm gene and 2) the region containing both MCM- and integrase-coding genes was flanked by direct repeats (attachment [att] sites). MCM-coding genes are referred to as cellular if they were embedded into the context of genes that are usually not present in viral genomes and/or plasmids, that is, responsible for energy production, cellular metabolism, and so forth.

Contextual Analysis of the MCM-Encoding Mobile Elements
Proteins encoded within the potentially mobile elements were analyzed using BlastP (Altschul et al. 1997) implemented within the CLC Main Workbench software package (CLC Bio, Inc.). The searches were done independently against the viral and whole nonredundant protein sequence databases at NCBI. The integrated elements were considered to be of viral origin (proviruses) if virion structural proteins were found to be encoded by the element.

Phylogenetic Analysis
MCM homologues were aligned using MUSCLE 3.6 (Edgar 2004), and the alignments were edited and refined manually using the ED program from the MUST package.
Results and Discussion

Nearly One-third of Archaeal MCM Proteins Are Encoded by Mobile Elements

Most archaeal genomes possess a single gene encoding an MCM helicase. However, some of them possess multiple mcm genes. These additional copies can arise from either gene duplications of the resident mcm or the integration of extrachromosomal elements containing mcm genes.

To elucidate the role of these two possible mechanisms in the origin of multiple mcm genes in some archaeal species, we set out to identify all sequenced MCM homologues associated with the domain Archaea. We identified 102 MCM homologues from 67 archaeal taxa and in one archaeal virus. Five of these 102 mcm genes were located within fragments sequenced during environmental sampling trials, and full genetic context for these mcm genes is therefore unavailable.

Thirteen archaeal species were found to contain more than one MCM-coding gene. We therefore inspected their genomes for signs of recombination signals proximal to the mcm genes (see Materials and Methods for details). Archaeal species often harbor integrated elements (plasmids and viruses) within their genome (Cortez et al. 2009). Integration of these elements occurs upon site-specific recombination between the invading element and the host chromosome and is usually catalyzed by the element-encoded integrases. Known archaeal integrases can be divided into two types based on the strategy of integration (She et al. 2004). In type I, recombination of the circular genetic element with the host chromosome leads to the partitioning of the integrase gene into longer Int (C) and shorter Int (N) fragments. Type II comprises enzymes that maintain an intact integrase-coding gene after recombination. In both cases, the att sites in the invading genetic elements are usually highly similar or identical to the 5′- or 3′-distal regions of tRNA genes, which are consequently used as the integration site (Reiter et al. 1989; Krupovic and Bamford 2008a). For type I integrases, the att site is located inside the viral integrase gene, whereas those of type II are often adjacent to the integrase gene (She et al. 2004). Upon integration of a circular genetic element, the att sequence flanks the inserted element as direct repeats (attL on the left side of the element and attR on the right). Our analysis showed that 23 mcm genes were encoded within genomic regions that most likely represent mobile elements integrated into the chromosome of their hosts. In addition, three archaeal plasmids and one virus were also found to encode MCM proteins (table 1). This suggests that 27 of 97 (27.8 %) archaeal MCM homologues for which genetic context information is available are encoded by mobile elements. Analysis of the complete genetic content suggested that 5 of the 23 integrated mobile elements are likely to be of viral origin (proviruses). Unfortunately, limited knowledge on archaeal viruses and plasmids does not allow us to firmly distinguish between the viral and plasmid origin for the rest of the identified elements. We will therefore generally refer to them as mobile elements. We carried out a detailed description of each of these elements.

Description of MCM-Encoding Mobile Elements

Tables presenting the annotation of all the identified integrated elements described in this study, as well as their precise nucleotide coordinates and att sequences, can be found as supporting information (see supplementary files 1 and 2, Supplementary Material online). The elements encode a number of different helicases, viral genome packaging ATPases, restriction–modification (R-M) systems, toxin/antitoxin systems, and transcriptional regulators but generally lack genes encoding proteins involved in metabolism or energy production. A substantial fraction of genes found within these potential mobile elements are also common to plasmids and viruses.

MCM-Encoding Plasmids

Only three plasmids encoding MCM proteins have been sequenced so far (table 1). pTAU4 of Sulfolobus neozelandicus is the only crenarchaeal plasmid that was reported to encode an MCM homologue. This mcm gene has been suggested to replace a gene coding for the primase/polymerase domain–containing RepA protein, which is otherwise present in Sulfolobus plasmids of the pRN family (Greve et al. 2005). Notably, none of the genomes from Crenarchaeota possess more than one identifiable chromosomal mcm gene copy. We found MCM homologues encoded by plasmids pNG300 and ECE1 (extrachromosomal element 1) of Haloarcula marismortui ATCC 43049 and Methanocaldococcus jannaschii DSM 2661, respectively. With the exception of the chromosome partitioning ParA-like Soj ATPase (COG1192) encoded by pNG300 (protein pNG3037) and ECE1 (protein MJECL24), MCM is the only protein common to all three S. neozelandicus, H. marismortui, and M. jannaschii plasmids. It is therefore likely that these three plasmids are unrelated and acquired their mcm genes independently.
as confirmed by the phylogenetic analysis of their MCM proteins (see below).

**MCM-Encoding (pro)Viruses**

Archaeal head-and-tail virus BJ1 infecting *Halorubrum* species encodes an MCM homologue (Pagaling et al. 2007). Interestingly, BJ1 also encodes a homologue of Cdc6/Orc1, which is involved in binding to the replication origin as well as in loading of the MCM helicase onto DNA (Matsumaga et al. 2001; Barry and Bell 2006), suggesting that these two virus-encoded proteins cooperate in the replication of the BJ1 genome. BJ1 is the only free-living archaeal virus isolated thus far that encodes an MCM helicase. However, search for viral genomes integrated into the chromosomes of diverse archaeal species revealed the presence of several proviruses encoding *mcm* genes (Krupovic and Bamford 2008a; Krupovic et al. 2010, table 1). These are integrated into the genomes of diverse archaeal species belonging to the taxonomic orders Halobacterialia, Methanosarcinales, Thermococcales, and Methanopyrales. Four of these MCM-encoding proviruses (Hlac-Pro1, Mace-Pro1, Mjan-Pro1, and MmarC6-E2) are clearly related to the head-and-tail bacterial and archaeal viruses of the order Caulovirales (Krupovic et al. 2010), whereas one provirus, TKV4, residing in the genome of *Thermococcus kodakarensis* KOD1, seems to be related to a tailless membrane-containing *Sulfolobus* turreted icosahedral virus (Krupovic and Bamford 2008a).

### Table 1. Archaeal MCM Proteins Encoded by Putative Mobile Elements.

<table>
<thead>
<tr>
<th>Order/Species Name</th>
<th>Element Name</th>
<th>Type of Element</th>
<th>att Site in the Host</th>
<th>MCM Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Sulfolobus</em> neozealandicus</td>
<td>pTAU4</td>
<td>Plasmid</td>
<td>—</td>
<td>CAH65772</td>
</tr>
<tr>
<td><strong>Halobacterialia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Natronomonas pharaonis</em> DSM 2160</td>
<td>Npha-E1</td>
<td>Element</td>
<td>tRNA-Thr</td>
<td>YP_327587</td>
</tr>
<tr>
<td><em>Haloarcula marismortui</em> ATCC 43049</td>
<td>Hmar-E1</td>
<td>Element</td>
<td>tRNA-Phe</td>
<td>YP_136848</td>
</tr>
<tr>
<td><em>H. marismortui</em> ATCC 43049</td>
<td>pNG300</td>
<td>Plasmid</td>
<td>—</td>
<td>YP_134081</td>
</tr>
<tr>
<td><em>Halorubrum lacusprofundi</em> ATCC 49239</td>
<td>Hlac-Pro1</td>
<td>Provirus</td>
<td>tRNA-Thr</td>
<td>YP_002565410</td>
</tr>
<tr>
<td><em>Halorubrum</em> species-infecting virus BJ1</td>
<td>—</td>
<td>Virus</td>
<td>—</td>
<td>YP_919062</td>
</tr>
<tr>
<td><strong>Methanosarcinales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Methanococcaea acticorans</em> C2A</td>
<td>Mace-Pro1</td>
<td>Provirus</td>
<td>intergenic</td>
<td>NP_618700</td>
</tr>
<tr>
<td><strong>Thermococcales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Thermococcus kodakarensis</em> KOD1</td>
<td>TKV1</td>
<td>Element</td>
<td>tRNA-Val</td>
<td>YP_182509</td>
</tr>
<tr>
<td><em>T. kodakarensis</em> KOD1</td>
<td>TKV4</td>
<td>Provirus</td>
<td>tRNA-Leu</td>
<td>YP_183774</td>
</tr>
<tr>
<td><strong>Methanopyrales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Methanopyrus kandleri</em> AV19</td>
<td>Mkan-E1</td>
<td>Element</td>
<td>N.D.</td>
<td>NP_614403</td>
</tr>
<tr>
<td><strong>Methanococcales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Methanocaldicoccus jannaschii</em> DSM 2661</td>
<td>ECE1</td>
<td>Plasmid</td>
<td>—</td>
<td>NP_044140</td>
</tr>
<tr>
<td><em>M. jannaschii</em> DSM 2661</td>
<td>Mjan-Pro1</td>
<td>Provirus</td>
<td>tRNA-Ser</td>
<td>NP_247337</td>
</tr>
<tr>
<td><em>Methanococcus aeolicus</em> Nankai-3</td>
<td>Maeo-E1</td>
<td>Element</td>
<td>tRNA-His</td>
<td>YP_001325414</td>
</tr>
<tr>
<td><em>Methanos. vannielii</em> SB</td>
<td>Mvan-E1</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001322788</td>
</tr>
<tr>
<td><em>Methanos. vannielii</em> SB</td>
<td>Mvan-E2</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001321444</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C5</td>
<td>MmarC5-E1</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001097170</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C5</td>
<td>MmarC5-E2</td>
<td>Element</td>
<td>N.D.</td>
<td>YP_001097634</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E1</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001548088</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E2</td>
<td>Provirus</td>
<td>tRNA-Ser</td>
<td>YP_001548164</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E3</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001548188</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E4</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001548471</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E5</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001548488</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C6</td>
<td>MmarC6-E6</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001548979</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C7</td>
<td>MmarC7-E1</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001329249</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> C7</td>
<td>MmarC7-E2</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>YP_001329286</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> S2</td>
<td>MmarS2-E1</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>NP_987590</td>
</tr>
<tr>
<td><em>Methanos. maripaludis</em> S2</td>
<td>MmarS2-E2</td>
<td>Element</td>
<td>tRNA-Ser</td>
<td>NP_987868</td>
</tr>
</tbody>
</table>

Note.—N.D., not determined.

**MCM-Encoding Mobile Elements**

Most of the mobile elements encoding MCM proteins, with the exception of the methanococcal ones (see below), are not identifiable related to each other. For example, the two halobacterial elements, about 14.5-kb-long Hmar-E1 in *H. marismortui* and about 56.4-kb-long Npha-E1 from *Natronomonas pharaonis* (table 1), share only the MCM and an integrase-coding genes, which are respectively 63% and 45% identical at the protein level (Supplementary fig. S1, Supplementary Material online). Interestingly, Npha-E1 encodes a set of five putative helicases in addition to the MCM helicase (Supplementary fig. S1A, Supplementary Material online). Similarly, the two MCM-encoding integrated elements of *T. kodakarensis* KOD1, TKV1 and provirus TKV4, do not share much similarity except for the related integrase and *mcm* genes (Fukui et al. 2005). Notably, all four integrated elements identified in the *T. kodakarensis* KOD1 genome (Fukui et al. 2005) use type I integrases (see above) for integration, whereas all other mobile elements described in this study rely on type II enzymes.
**Methanopyrus kandleri** AV19 is the only member of the order Methanopyrales for which a complete genome sequence is currently available (Slesarev et al. 2002). Previous analysis of the *Methanopyrus kandleri* genome by Jensen et al. (2003) revealed two enigmatic regions (regions I and II) containing a large fraction of genes coding for orphan proteins with no known homologues in other organisms (Jensen et al. 2003, Supplementary fig. S2, Supplementary Material online). The authors hypothesized that the two regions represent integrated plasmids, although putative att sites could not be identified. Interestingly, we found that one of the two *Methanopyrus kandleri* mcm genes, mcm-2, is within region I (denoted here as Mkan-E1; Supplementary fig. S2, Supplementary Material online). The structure of *Methanopyrus kandleri* MCM-2 has been solved by X-ray crystallography (Bae et al. 2009). The authors screened two dozen constructs of various archaeal MCM homologues (including those from *Archaeoglobus fulgidus*, * Aeropyrum pernix*, *Methanosarcina acetivorans*, and *Methanopyrus kandleri*), but only MCM-2 from *Methanopyrus kandleri* yielded diffraction-quality crystals. Particularly, the sequence of MCM-2 lacks the amino-terminal zinc-binding motif as well as the C-terminal HTH motif and harbors non-canonical residues at the AAA ATPase active site. Furthermore, recombinant MCM-2 is monomeric in solution (not hexameric) and does not exhibit helicase activity (Bae et al. 2009). These results, together with our observation that MCM-2 originates from a putative mobile element, suggest that this protein is inactive and is unlikely to be involved in replication of the *Methanopyrus kandleri* chromosome, a function likely performed by MCM-1.

There are currently eight complete genome sequences available for members of the archaeal order Methanococcales. Strikingly, all of them encode more than two MCM homologues, except *Methanococcus voltae* A3, which encodes two. In all cases, two mcm genes are embedded within a cellular gene context and have likely originated due to an ancient gene duplication event (see also Phylogenetic Analysis of Archaeal MCM Helicases below). *Methanococcus maripaludis* C6 encodes a record number of MCM homologues on its circular chromosome. In addition to the two cellular mcm genes, it possesses six mcm genes that are associated with potential mobile elements, MmarC6-E1–6 (figs. 1 and 2 and table 1). Such a high number of MCM-encoding elements in members of the order Methanococcales, which is unseen in other archaeal species, warrants a more detailed description of these elements (see below).

**MCM-Encoding Mobile Elements in Members of Methanococcales**

Elements MmarC6-E1, -E2, and -E3 are adjacent to each other, separated by identical att sites (see supplementary material S1, Supplementary Material online). The putative att sequence used by these elements is identical to the 3’-distal region of a tRNA\(^\text{Ser}\) gene (fig. 1). The simplest explanation for the origin of this region of concatenated mobile elements is that each of the elements recombined with the target tRNA gene independently. It should be noted that upon integration of an element into a target tRNA gene, the original gene sequence is preserved. Therefore, the first element most likely integrated into the element-free tRNA gene, whereas the consecutive recombinations might have occurred either with the element-containing tRNA gene or with the direct repeat flanking the element on the other side than the tRNA gene. A similar integration of two different proviruses into the same tRNA gene has been previously reported in *Methanococcus voltae* A3 (Krupovic and Bamford 2008a). Interestingly, MmarC6-E6 uses the same tRNA\(^\text{Ser}\) gene for integration as MmarC6-E1–3. However, the att sequence of MmarC6-E6 is identical to the 5’- but not to the 3’-distal region of the tRNA gene (fig. 1). Integrative plasmids and viruses usually recombine with the 3’-distal regions of their target tRNA genes (Reiter et al. 1989), but recombination with the 5’-distal region of a tRNA gene has been reported for provirus TKV4 of *T. kodakarenensis* KOD1 (Krupovic and Bamford 2008a). MmarC6-E4 and -E5 are also adjacent to each other, separated by att sites that are identical to the 3’-distal region of another tRNA\(^\text{Ser}\) gene (fig. 1).

The six elements of *Methanococcus maripaludis* C6 can be categorized into four groups. Elements MmarC6-E1, -E2, and -E6 form groups of their own, whereas MmarC6-E3, -E4, and -E5 are related to each other (fig. 2, also see below). MmarC6-E1 is the largest (48.2 kb) of the six elements. Its 5’-distal region, encompassing eight genes, is significantly similar and collinear with regions of integrated elements TKV2 and TKV3 of *T. kodakarenensis* KOD1 (Fukui et al. 2005), and plasmid pT26-2 from *Thermococcus* species 26.2 (Keller et al. 2009; Soler et al. 2010; see the corresponding annotation table for more details, supplementary file 2, Supplementary Material online). MmarC6-E2 encodes a number of proteins that share significant sequence similarity with those of head-and-tail viruses and related proviruses (Krupovic et al. 2010), suggesting that MmarC6-E2 represents an integrated genome of a tailed archaeal virus (fig. 2). MmarC6-E3, -E4, and -E5 share a core of eight genes, including those coding for an MCM helicase, an integrase, and a PadR/MarR-like transcriptional regulator (Pfam: PF03551). MmarC6-E3 and -E5 encode four and MmarC6-E4 encodes three proteins that are also found to be encoded by 8,285-bp-long plasmids pURB500 (GenBank accession no.: U47023) and pMMC501 (GenBank accession no.: CP000610) of *Methanococcus maripaludis* C5 (see corresponding annotation tables, supplementary file 2, Supplementary Material online), suggesting a possible evolutionary relationship between these elements.

MmarC6-E2 is distantly related to the Mjan-Pro1 provirus of *M. jannaschii* DSM 2661 and is otherwise not found in other members of the Methanococcales (Krupovic et al. 2010). In contrast, elements related to MmarC6-E1 are present in all *Methanococcus maripaludis* strains sequenced so far (supplementary table S1 and fig. S3, Supplementary Material online). Upon integration, all E1-like elements apparently recombined with the 3’-distal regions of tRNA\(^\text{Ser}\) genes. Comparative analysis revealed a number of deletions/insertions and rearrangements within this
group of elements (Supplementary fig. S3, Supplementary Material online). For example, E1-like element of *Methanoc. maripaludis* S2 encodes an McrBC restriction enzyme, whereas E1-like elements in the other three strains of *Methanoc. maripaludis* encode a complete set of enzymes constituting type I R-M systems (Supplementary fig. S3, Supplementary Material online). E1-like elements of *Methanoc. maripaludis* strains C6, C7, and S2, but not C5, possess TKV2/3-like regions (see above). MmarC7-E1 of *Methanoc. maripaludis* C7 is nearly twice in size compared with MmarC6-E1. Half of the MmarC7-E1 element is highly similar to the corresponding elements of *Methanoc. maripaludis* C6 and S2 both in gene content and in arrangement (Supplementary fig. S3, Supplementary Material online). However, the other half of the element contains genes that have no known viral or plasmid homologues but are abundant in cellular genomes. These genes encode proteins such as dehydrogenase maturation factor (COG364), lysine 2,3-aminomutase (COG1509), acetyltransferase (COG3153), flavin-nucleotide-binding protein (COG3467), and multidrug efflux pump (COG0534; complete annotation table of MmarC7-E1 can be found in supplementary file 2, Supplementary Material online). A likely explanation for the origin of the MmarC7-E1 element is that recombination between the possibly preintegrated ancestral element and the cellular chromosome resulted in the insertion of the chromosomal region between the two att sites of the mobile element. Interestingly, MmarC5-E1 is a chimera of MmarC6-E1 and MmarC6-E6. Two-thirds of the element constitute of genes common to E1-like elements from the other *Methanoc. maripaludis* strains (see the corresponding annotation table, supplementary file 2, Supplementary Material online), whereas the 3′-distal region of the element contains genes, including the one encoding the MCM helicase, that are similar to those found in the MmarC6-E6 element (Supplementary fig. S3, Supplementary Material online).

Elements related to the MmarC6-E3/E4/E5 are also present in the genomes of *Methanoc. maripaludis* strains C5 and S2 (Supplementary fig. S4, Supplementary Material online). Each of the two strains contains a single element of this type (supplementary table S1, Supplementary Material online). Interestingly, the MmarC5-E2 of *Methanoc. maripaludis* C5 is highly unlikely to be still mobile as we were unable to identify the putative att sites, nor was the
The integrase gene present at the position equivalent to those in other related elements (Supplementary fig. S4, Supplementary Material online). However, on the right side of the element we identified a sequence that is identical to the att sequence of the MmarS2-E1 (see Supplementary Material for more details). It is therefore plausible that the ancestor of MmarC5-E2 integrated into the target sequence equivalent to the one used by MmarS2-E1 in Methanoc. maripaludis S2, and was later inactivated by the loss of one of the att sequences along with the integrase gene. Interestingly, of the four MCM proteins of Methanoc. maripaludis S2, at least three have been recently shown to be expressed (Walters and Chong 2009). The whole proteome of Methanoc. maripaludis S2 cells grown in nitrogen-free medium has been analyzed by Xia et al. (2006) using quantitative multidimensional capillary high-performance liquid chromatography, followed by quadrupole ion-trap mass spectrometry. Peptides were identified for both cellular MCMs and the one encoded within the MmarS2-E1 element. No peptides were detected for the MCM encoded by the MmarS2-E2. This observation suggests that mobile element–encoded MCMs in Methanoc. maripaludis S2, and possibly in other archaeal species as well, might not be just silent genome “passengers” but rather might contribute (either positively or negatively) to the replication of the cellular genome.

In addition to the chimeric element MmarC5-E1, MmarC6-E6-like elements are present in Methanoc. maripaludis C7 (MmarC7-E2) and Methanoc. vannielii SB (Mvan-E1; Supplementary fig. S5 and table S1, Supplementary Material online). MmarC7-E1 and -E2 of Methanoc. maripaludis C7 are nearly identical to MmarC6-E1 and -E6 of Methanoc. maripaludis C6 (fig. 1), MmarC7-E1 and -E2 of Methanoc. maripaludis C7 are integrated into the same tRNA^Ser_ gene where they have recombined with the 3′- and 5′-distal regions of the...
Maeo-E1 is integrated into a gene encoding a tRNAHis, differently from all the other methanococcal elements, encoded by TKV4 from Methanococcus vannielii (GenBank accession no.: YP_001324144). This fragmented mcm gene is within a 14.6-kb-long element Mvan-E2, which is integrated into a tRNA^Ser^ gene (Supplementary fig. S6A, Supplementary Material online).

*Methanococcus aeolicus* Nankai-3 genome contains three genes coding for MCM helicases. Two of these are apparently cellular, whereas the third one is within a 15.2-kb-long potential mobile element (Maeo-E1; Supplementary fig. S6B, Supplementary Material online). Differently from all the other methanococcal elements, Maeo-E1 is integrated into a gene encoding a tRNA^His^, not tRNA^Ser^. Most of the genes within this element have no known homologues (see the corresponding annotation table, supplementary file 2, Supplementary Material online). Peculiarly, Maeo-E1 encodes a putative ATPase homologous to the FtsK/HerA-like packaging ATPases of membrane-containing viruses (Iyer et al. 2004; Strömssten et al. 2005; Krupovic and Bamford 2008b), suggesting that Maeo-E1 might originate from the integration of a virus.

**Phylogenetic Analysis of Archaeal MCM Helicases**

To understand the evolutionary history of the various archaeal MCM proteins analyzed in this study, we performed Bayesian phylogenetic analysis (maximum likelihood analysis gave very similar results; data not shown). The resulting tree (fig. 3) shows, with the exception of the MCM protein encoded by TKV4 from *T. kodakarenensis*, all MCMs belonging to mobile elements cluster with the cellular MCMs of the host group (e.g., Halobacteriales, Methanosarcinales, Thermococcales, Sulfolobales, Methanococcales), indicating that they share a common origin with MCM of their hosts. It should be noted that all multiple MCM copies from Methanococcales cluster together, indicating a common origin. MCM belonging to plasmids, viruses, or integrated mobile elements typically exhibit longer branches compared with the closely related cellular copies. This indicates that these MCM proteins have experienced an accelerated rate of protein evolution after their recruitment by mobile elements. As a consequence, their position as outgroups of their cellular relatives can reflect either an ancient origin (before the divergence of the group) or be due to the long-branch artifact produced by differences in evolutionary rates. The placement of some “noncellular” MCMs was indeed very unstable due to their high divergence. This is, for example, the case for the proviral MCM (in TKV4) from *T. kodakarenensis*, which branched at the base of Methanococcales (fig. 3). In that case, this placement is clearly an artifact because this MCM clusters with Thermococcales as well as with the second “mobile” MCM (encoded by TKV1) from *T. kodakarenensis* in a tree built from a data set where Methanococcales sequences were removed (data not shown). In general, the placement of mobile MCMs was strongly dependent on the taxonomic sampling, whereas this was not observed with cellular sequences. This suggests that caution should always be taken in interpreting trees including highly diverging sequences such as the viral ones.

Indeed, when we built a tree using only cellular MCM sequences (fig. 4), this tree was strikingly congruent with the current consensus on archaeal phylogeny (Brochier-Armanet et al. 2008; Gribaldo and Brochier 2009) where all the major phyla and internal orders were recovered. This means that an mcm gene was present in the ancestor of the Archaea and that these genes have not been affected by HGT among different lineages during the evolutionary history of this domain. Interestingly, the recently proposed phylum Thaumarchaeota (Brochier-Armanet et al. 2008) is firmly placed at the base of the archaeal tree, in line with the result recently obtained with a concatenation of ribosomal proteins (Brochier-Armanet et al. 2008). Our analysis thus suggests that at least some DNA replication proteins can be probably added to the set of proteins used to reconstruct archaean phylogeny and confirms the existence of a core of genes that can be used to reconstruct the history of the archaean domain (Gribaldo and Brochier 2009).

Although the removal of MCM proteins located within mobile element was necessary to obtain a phylogeny completely congruent with the consensus archaean phylogeny, it is remarkable that the complete MCM tree including both cellular and mobile MCMs is still roughly congruent with the consensus archaean phylogeny. This indicates that mobile elements carrying mcm genes have mainly co-evolved with their hosts and were not involved in extensive lateral gene transfer between evolutionary distantly related archaean groups. This is in agreement with recent data on the evolution of several new plasmids isolated from Thermococcales (Soler et al. 2010) but somehow in contradiction with current view that viruses and plasmids promote extensive exchanges between evolutionary distantly related cellular lineages (Moreira and Lopez-Garcia 2009).

**Evolutionary History of MCMs in Methanococcales**

The very long branch leading to the MCMs from Methanococcales, the occurrence of a gene duplication in the ancestor of the group, and the presence of multiple *mcm* genes embedded in mobile elements are all intriguing characteristics. Moreover, it has been recently shown that MCMs from Methanococcales, both cellular and element encoded, form heterocomplexes in vitro (Walters and Chong 2010). This prompted us to investigate the evolutionary relationships of these cellular- and element-associated MCMs in more detail. We therefore performed a separate phylogenetic analysis restricted to MCM sequences of Methanococcales to have more amino acid positions to analyze and to limit possible artifacts introduced by the...
Bayesian phylogenetic tree of archaeal MCMs rooted by a selection of eukaryotic homologues as outgroup. After removal of ambiguously aligned regions, the final data set contained 343 amino acid positions for analysis. The different types of mobile elements are indicated by different colors. The scale bar represents the average number of substitutions per site. Numbers at nodes represent confidence intervals.
presence of other fast-evolving sequences (fig. 5). Apart from the already mentioned cellular MCM groups I and II, the MCM homologues from integrated elements clearly segregate into two groups (III and IV). Group III branches within cellular group II and likely originated from acquisition by mobile elements of one of the two paralogous
cellular mcm copies from the ancestor of *Methanoc. vannielii* and *Methanoc. maripaludis* (red arrow in fig. 5). This is consistent with the presence of group III MCM in *Methanoc. vannielii* and *Methanoc. maripaludis* only. Interestingly, group III MCMs subdivide into two subgroups (IIa and IIIb). Group IIIb corresponds to E3/4/5-like elements (fig. 2 and Supplementary fig. S4, Supplementary Material online), whereas group IIa is more diverse and includes E1-
like elements as well as the MmarC6-E2 provirus of Methanoc. maripaludis C6 (fig. 2 and Supplementary fig. S3, Supplementary Material online). It is possible that these two groups of elements acquired their mcm genes independently. For example, the MCM from the E2-like provirus of M. jannaschii (Krupovic et al. 2010) appears to be only distantly related to the MCM of the MmarC6-E2 element from Methanoc. maripaludis C6 (fig. 5). Group IV includes E6-like elements (Supplementary fig. S5, Supplementary Material online) and branches in between the two cellular MCM groups. A recent phylogenetic analysis of Methanococcales MCM has led to the proposal that MCMs belonging to group IV originated from duplication of cellular group II mcm (Walters and Chong 2010). However, we believe that two hypotheses are possible for the origin of group IV MCMs: 1) either group IV MCM also originated from cellular ones but their fast evolution leads to misplacement in the tree or 2) they may have been acquired by an ancestral mobile element before the diversification of Methanococcales and coevolved with Methanococcales within a viral/plasmid world; some of these elements would have then been integrated into cellular chromosomes in Methanoc. maripaludis and Methanoc. vanneili.

The long branch at the base of Methanococcales MCM, along with the unique ancient duplication and abundance of mobile elements, is intriguing. Such fast evolution of MCMs originating from Methanococcales may be linked to the duplication that produced two MCM copies that adapted to function as a heterocomplex (Walters and Chong 2010). Mobile elements may have taken advantage of such peculiarity of Methanococcales to recruit MCM to highjack the host replication system by forming heterocomplexes with cellular MCM, although this remains to be proven in vivo. In contrast, the long branch at the base of Methanococcales in the cellular tree strongly reminds the long branches observed for all mobile MCMs, suggesting an alternative scenario in which 1) a virus infecting the ancestor of Methanococcales acquired an mcm gene from its host; 2) this mcm gene experienced fast evolution within a viral/plasmid world; 3) it was reintroduced into the host chromosome of the ancestor of Methanococcales, replacing the native mcm copy; (iv) the mcm gene was duplicated leading to the two cellular MCM copies; and (v) over subsequent evolution of Methanococcales, there were further exchanges between the host and the mobile elements. According to this scenario, what now appears to be the cellular mcm might indeed represent duplication of an anciently introduced viral/plasmid mcm that has replaced the original host gene.

Interestingly, Methanococcales are the only archaea (with the exception of Methanop. kandleri) that harbor no clear-cut homologues of the eukaryotic initiator proteins Cdc6/Orc1, used to initiate DNA replication in other Archaea (Zhang and Zhang 2004; Berthon et al. 2008). Because MCM and Cdc6/Orc1 functionally and physically interact to initiate DNA replication (Barry and Bell 2006), the loss of Cdc6/Orc1 might have relaxed the evolution of MCM in Methanococcales. A gene encoding a putative initiator protein (MJ0774) has been detected near the predicted replication origin of M. jannaschii (Zhang and Zhang 2004). This protein harbors the AAA + ATPases domain, as is the case for other initiator proteins, such as Cdc6/Orc1 and DnaA. Homologues of this protein are present in all Methanococcales but also in some Thermococcales. This protein might correspond to the initiator protein of a viral/plasmid family that coevolves with Methanococcales and Thermococcales, and have replaced the canonical archaeal Cdc6/Orc1 as a cellular DNA replication initiator protein at the onset of the Methanococcales lineage. It is tempting to speculate that this protein have been transferred together with the viral MCM protein encoded by the same mobile element or that it recruited the MCM proteins from another mobile element instead of the cellular MCM to perform its new task as a cellular initiator. However, these scenarios cannot be validated because MJ0774 and its archaeal homologues are presently located within cellular gene context.

Further Insights into Methanococcal MCM Helicases

Presence of multiple MCM copies in Methanococcales a priori suggests that the active helicase in these organisms might be heteromultimeric rather than homomultimeric. This possibility is especially attractive because MCMs encoded in mobile elements might form heteromultimeric complexes with cellular MCMs to specifically replicate their genomes. To better understand the specificities of MCMs from Methanococcales, as well as their evolutionary and functional relationship, especially between the cellular and mobile ones, we extracted the multiple sequence alignment of the four major groups (G1–4) of methanococcal MCMs from the global alignment along with some representatives of other archaean MCMs for more comprehensive examination. The methanococcal sequences were analyzed in the framework of the structural information available for MCMs from Methanothermobacter thermautotrophicus (MthMCM; PDB ID: 1LTL) and S. solfataricus (SsoMCM; PDB ID: 3F9V) (Supplementary fig. S7, Supplementary Material online).

Archaean MCM protein can be structurally subdivided into three major domains: the N-terminal, the central AAA+ ATPase, and the C-terminal ones (fig. 6A). The N-terminal domain can be further subdivided into three subdomains, sA, sB, and sC, whereas the central AAA+ ATPase domain consists of subdomains α/β and α (Fletcher et al. 2003; Brewster et al. 2008). The C-terminal domain is predicted to adopt the wHTH fold (Aravind and Koonin 1999). Available MCM structures as well as functions of the structural and regulatory elements of these proteins have been recently reviewed (Bochman and Schwacha 2009; Costa and Onesti 2009; Sakakibara et al. 2009). We will therefore only discuss the ones that are relevant to our analysis.

Subdomains sB and sC play an important role in the MCM hexamer formation (Fletcher et al. 2003; Liu et al.
Insights into the structure of methanococcal MCM helicases. (A) Domain organization of the four groups of methanococcal MCM helicases (G1–4) in comparison with that of *Sulfolobus solfataricus* (Sso). Typical MCM helicase consists of three domains: N-terminal, central AAA+ ATPase, and C-terminal. The N-terminal domain is subdivided into three subdomains: sA (cyan), sB (magenta), and sC (green). The AAA+ ATPase domain is subdivided into α/β (red) and α (orange) subdomains that are connected via the α/β-α linker (light gray). The N-terminal and the AAA+ ATPase domains are connected through the N–C linker (dark gray). The C-terminal domain (dark blue) is predicted to adopt the wHTH fold. The functional elements are also indicated: Zn finger, zinc finger; ACL, allosteric communication loop; N-t hp; EXT hp; W-A; H2i; W-B; PS1 hp, presensor 1 β hairpin; S1; RF; PS2; and S2. Group-specific insertions and deletions within the methanococcal MCMs are indicated with black triangles pointing down and up, respectively. Asterisk denotes that the N-t hp in methanococcal MCMs is shorter than in *Sulfolobus solfataricus*.

**Fig. 6.** Insights into the structure of methanococcal MCM helicases. (A) Domain organization of the four groups of methanococcal MCM helicases (G1–4) in comparison with that of *Sulfolobus solfataricus* (Sso). Typical MCM helicase consists of three domains: N-terminal, central AAA+ ATPase, and C-terminal. The N-terminal domain is subdivided into three subdomains: sA (cyan), sB (magenta), and sC (green). The AAA+ ATPase domain is subdivided into α/β (red) and α (orange) subdomains that are connected via the α/β-α linker (light gray). The N-terminal and the AAA+ ATPase domains are connected through the N–C linker (dark gray). The C-terminal domain (dark blue) is predicted to adopt the wHTH fold. The functional elements are also indicated: Zn finger, zinc finger; ACL, allosteric communication loop; N-t hp; EXT hp; W-A; H2i; W-B; PS1 hp, presensor 1 β hairpin; S1; RF; PS2; and S2. Group-specific insertions and deletions within the methanococcal MCMs are indicated with black triangles pointing down and up, respectively. Asterisk denotes that the N-t hp in methanococcal MCMs is shorter than in
Furthermore, subdomain sC in MthMCM was shown to be responsible for the interaction with the Cdc6/Orc1 proteins (Kasiviswanathan et al. 2005). Methanococcal G2 (cellular) and G3 MCMs (element) have a specific and conserved seven to eight-amino-acid insertion within the β3 strand of the subdomain sC (here and later on the secondary structure element nomenclature is as shown for the SsoMCM in Supplementary fig. S7, Supplementary Material online). In addition, there is a deletion of two amino acids between strands β10 and β11 of the sC subdomain that is specific to Methanococcales (Supplementary fig. S7, Supplementary Material online), to the exception of the MCM encoded by the putative provirus TKV4 of T. kodakarenensis KOD1 (data not shown). These modifications can be rationalized by the fact that the Methanococcales MCMs do not interact with Cdc6/Orc1 proteins (missing in these species) but with another initiator protein.

Notably, the N-terminal β-hairpin (N-t hp), which resides in the central channel of the hexameric MCM helicase and is involved in DNA binding, is shorter in MCMs from Methanococcales when compared with that of S. solfataricus and resembles the N-t hp of Methanococcus thermotrophicus (Fletcher et al. 2003; Liu et al. 2008). Consequently, the diameter of the central channel is expected to be similar to that of MthMCM (Fletcher et al. 2003). The sB subdomain contains the zinc finger motif, which is also involved in DNA binding (Sakakibara et al. 2009). The motif is present in all methanococcal MCMs. However, the sB subdomain in the G1 (cellular) MCM of M. jannaschii displays a number of insertions that are not observed in other sequences of G1 (Walters and Chong 2010, Supplementary fig. S7, Supplementary Material online).

The N–C linkers in G1 and G4 MCMs are longer than in other methanococcal MCMs as well as those of MthMCM and SsoMCM due to two independent five-amino-acid insertions in the N-terminal to α6 and α7, respectively (Supplementary fig. S7, Supplementary Material online). The α/β subdomain is the best-conserved region of archaeal MCM helicases. Nonetheless, the G2, G3, and G4 MCMs harbor a deletion of five amino acids within the external β-hairpin (EXT hp), with G4 MCM having a deletion of an additional amino acid, which is well conserved in all archaean MCM (fig. 6A and Supplementary fig. S7, Supplementary Material online). The EXT hp is well conserved in archaean MCMs and was shown to be critical for helicase activity (Brewster et al. 2008). The deletions in this region in methanococcal G2–4 MCMs is therefore surprising and suggests that either EXT hp is not uniformly important for helicase activity in different archaean MCMs or that the primary role of MCMs with deletions within this loop is other than DNA unwinding per se but perhaps modulation of the activity specificity of the catalytically functional MCM copy within a potentially heteromultimeric MCM assembly.

Another notable characteristic of methanococcal MCMs when compared with the characterized MthMCM and SsoMCM is an insertion of two (in G1) or three (in G2–4) amino acids within the helix-2 insert hairpin (H2i; fig. 6A). The H2i is located in the central channel and was shown to be crucial for DNA unwinding activity of the MthMCM and was suggested to act as a molecular plough for mechanical separation of the two antiparallel strands of the duplex DNA (Jenkinson and Chong 2006; Brewster et al. 2008). An insertion within the H2i might therefore have an impact on the activity of the methanococcal MCMs.

The hydrolysis of ATP by MCM helicases is accomplished by a concerted action of Walker A (W-A), Walker B (W-B), sensor 1 (S1), arginine finger (RF), and sensor 2 (S2) elements (Moreau et al. 2007; Brewster et al. 2008, fig. 6A). In most multimeric AAA+ ATPases, W-A, W-B, S1, and S2 are provided in cis, whereas RF is acting in trans. Presensor 2 (PS2) α-helical insertion in MCM proteins enables the S2 to act in trans together with the RF (Erzberger and Berger 2006; Moreau et al. 2007; Brewster et al. 2008). The length of the PS2 insertion is generally uniform among archaean MCM (data not shown). However, the PS2 of G1 methanococcal MCMs is an exception. These MCMs have a specific insertion of about 20 amino acids, which is not observed in other archaeal MCMs (fig. 6A and Supplementary fig. S7, Supplementary Material online). As recently noticed (Walters and Chong 2010), this is reminiscent of the PS2 element found in eukaryotic MCMs, which also show differences among MCM families (Bochman and Schwacha 2009). Interestingly, G2–4 (but not G1) methanococcal MCMs have an approximately five-amino-acid deletion...
before the S2 element (Supplementary fig. S7, Supplementary Material online). This deletion may affect the function by altering the position of the essential S2 region with respect to other elements involved in ATP hydrolysis and helicase activity.

The predicted wHTH domain present at the extreme C-terminal end of archaeal MCM proteins has been suggested to modulate ATP hydrolysis, DNA binding, and duplex unwinding (Aravind and Koonin 1999; Jenkinson and Chong 2006). Surprisingly, methanococcal G4 MCMs lack the C-terminal region containing the wHTH domain (fig. 6A). Notably, the G4 mcm gene of Methanoc. vannielii SB contains a frameshift mutation that results in the loss of the C-terminal domain starting with the z subdomain. In silico correction of the reading frame allowed reconstitution of the z subdomain sequence, but the region encompassing the wHTH motif was not identified (data not shown). This indicates that the absence of the C-terminal wHTH motif is specific to all G4 MCMs. Particularly, the G4 MCMs have N-terminal extensions that are not observed in other archaeal MCMs (fig. 6). These extensions are approximately 90 amino acids in MCMs of Methanoc. maripaludis C6 (YP_001549879), Methanoc. maripaludis C7 (YP_001329286), and Methanoc. vannielii SB (YP_001322788), and approximately 50 amino acids in the MCM of Methanoc. maripaludis C5 (YP_001097170). PSI-Blast searches with these extensions as seeds did not reveal any close homologues. We therefore used the HHpred tool for distant homology detection (Söding 2005). Interestingly, HHpred analysis revealed a region with similarity to numerous wHTH motif-containing proteins for both the longer extensions (amino acids 54–93 of MCMs from Methanoc. maripaludis C6 and C7; prediction probability 68.53%) and the shorter one (amino acids 6–40 of MCM from Methanoc. maripaludis C5; prediction probability 63.23%). PsiPred secondary structure prediction (Bryson et al. 2005) for the N-terminal extension was also consistent with the presence of a wHTH fold. Figure 6B shows a sequence alignment of the region within the N-terminus of G4 MCMs with the C-terminal wHTH motifs of the representative G1–3 methanococcal MCMs and a selection of homologues identified by HHpred.

To get further insight into the topology of the N-terminal extension of the G4 MCMs, we submitted its sequence (100 amino acids of the Methanoc. maripaludis C6 MCM) to the I-TASSER server (Zhang 2008) for structure prediction. I-TASSER uses ab initio as well as homology-based approaches for structural modeling. The top model generated by I-TASSER was then used as a seed in the Vector Alignment Search Tool for structural neighbors (Gibrat et al. 1996). More than two hundred structures were identified. A common feature of the proteins related to the structural model of the N-terminal region of the G4 methanococcal MCMs was the presence of a wHTH motif. The top three structures were 1) the DNA-binding domain of the Escherichia coli ferric uptake regulator (PDB ID: 2FU4; 70 aligned residues), 2) the DNA-binding protein from M. jannaschii (PDB ID: 1KU9; 63 aligned residues), and 3) the DNA-binding domain of the transcription factor E from S. solfataricus (PDB ID: 1Q1H; 60 aligned residues). Structural alignment of the generated model with the three X-ray structures (fig. 6C) immediately reveals the wHTH motif in all four proteins. The wHTH of G4 MCMs if formed by α2x3β1β2, where β1 and β2 form the “wing,” whereas α3 is the DNA recognition helix that makes contact with the major groove of the DNA double helix, consistent with other wHTH proteins (Aravind and Koonin 1999). Notably, all G4 MCMs are encoded by mobile elements. It is therefore tempting to speculate that the loss of the wHTH from the C-terminus and acquisition of a similar motif at the N-terminus was selected for and plays a role in the replication of the mobile elements, possibly by modulating the activity and/or specificity of the host replication machinery. It should also be noted that there is no sequence similarity between the N-terminal region of G4 MCMs and the C-terminal regions of other archaeal MCM helicases, suggesting that the N-terminal wHTH motif of G4 MCMs did not originate from the transfer of the C-domain but was rather recruited from a different source.

To sum up, one insertion is common to G2 and G3, two deletions are common to G2–4, and an insertion is specific to G1 (fig. 6A). These characteristics correlate with the phylogeny of Methanococcales MCM: G2 and G3 are closely related groups, whereas G4 is more related to G2 and G3 than to G1, but has recently acquired its specific N-terminal wHTH domain and lost its C-terminal wHTH domain. The presence of this domain at a different location may indicate a peculiar regulation of the activity of G4 MCM. The deletion of the external hairpin, as well as deletion preceding S2 element in G2–4, could suggest that these MCMs might be functional only when associated with G1 as heteromers.

**Conclusions**

Several archaeal genomes encode multiple copies of the essential DNA replication protein, the MCM helicase. Our analysis indicates that most of these multiple copies did not arise from gene duplication but from the integration of mcm genes located within mobile elements. It has been suggested that paralogous MCM proteins in Archaea could be interesting models for eukaryotic MCM function because the eukaryotic functional helicase is a complex of six paralogous MCM proteins (Walters and Chong 2009). Our analysis clearly indicates that multiplication of mcm genes occurred independently in Archaea and Eukarya. However, it is possible that the multiple mcm genes that were already present in the last common eukaryotic ancestors originated from independent integration of mobile MCMs, as in the case of Archaea, and not by gene duplication, as usually believed (Wolf et al. 2004). Some MCMs encoded by mobile elements are now structurally distinct from their cellular counterparts (accumulation of insertion/deletion, domain swapping). In many cases, these MCMs are probably functional because they are present in free plasmids or viruses, or else conserved in divergent species (in Methanococcales). Structural modifications of
MCMs encoded within mobile elements could be related to their interaction with specific viral/plasmid DNA replication proteins (especially initiator proteins). In the case of Methanococcales, such modifications can be correlated with the lack of the Cdc6/Orc1 protein in these species and with the necessity for MCM to interact with a different initiator protein. The absence of the typical Cdc6/Orc1 protein in Methanococcales could also explain the complex evolution of MCM in this archaeal group, including gene duplication and the probable replacement of the cellular MCM by the MCM of a mobile element. The variety of MCMs in mobile elements compared with those encoded by cellular genomes illustrates the capacity of viruses and plasmids to produce new forms of DNA replication proteins.

Our analysis shows that MCM genes within mobile elements have been recruited relatively recently from cellular genomes (after the divergence of the various archaeal orders) and probably replaced DNA replication proteins specific for mobile elements, as in the case of pRN plasmids (Greve et al. 2005). Notably, mcm genes are not the only genome replication–associated genes that seem to have been recruited by archaeal mobile elements from their hosts. For example, different archaeal viruses, proviruses, and plasmids are known to encode homologues of the origin binding complex Cdc6/Orc1 (Yamashiro et al. 2006; Pagaling et al. 2007; Krupovic et al. 2010), whereas the integrated element TKV3 of T. kodakaraensis KOD1 encodes a homologue of a DNA polymerase sliding clamp (Fukui et al. 2005). Interestingly, despite extensive transfer of mcm genes between cells and mobile elements, the archaeal phylogeny obtained with MCM proteins is strikingly congruent with those obtained previously with ribosomal proteins or RNA polymerase subunits (Brochier-Armanet et al. 2008). The few differences observed can be corrected by removing from the analysis MCMs encoded within mobile elements. This indicates that MCM can be added to the list of core proteins useful to reconstruct the history of the archaeal domain. Proteins with multiple paralogs are usually systematically excluded from protein data sets used to reconstruct the history of species. We suspect that many cases of hidden paralogy are in fact due to the integration of mobile elements. We predict that many more proteins could be included in these data sets by systematically identifying and removing proteins encoded by mobile elements.

Supplementary Material
Supplementary files 1 and 2, figures S1–7, and table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
This work was supported by the Finnish Center of Excellence Program (2006–2011) of the Academy of Finland (grant 1129684 to D.H.B.) as well as the ARC (Association de la recherche contre le cancer) grant to P.F. M.K. was supported by a Federation of European Microbiological Societies research grant.

References


