Article

Not peer-reviewed version

Inteins at Eleven Distinct Insertion Sites in Archaeal Helicase Subunit MCM Exhibit Varied Architectures and Activity Levels Across Archaeal Groups

Danielle Arsenault *, Gabrielle Stack, Johann Peter Gogarten *

Posted Date: 28 May 2025

doi: 10.20944/preprints202505.2186.v1

Keywords: inteins; multi-intein genes; archaea; replicative DNA helicase

 doi:10.20944/preprints202505.2186.v1

Article

Inteins at Eleven Distinct Insertion Sites in Archaeal Helicase Subunit MCM Exhibit Varied Architectures and Activity Levels Across Archaeal Groups

Danielle Arsenault [1,*], Gabrielle F. Stack [1] and Johann Peter Gogarten [1,2,*]

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA

[2] Institute for Systems Genomics, University of Connecticut, Storrs, CT 06268-3125, USA

* Correspondence: danielle.arsenault@uconn.edu; gogarten@uconn.edu

Abstract: Background/Objectives: Inteins are mobile genetic elements invading highly conserved genes across all domains of life and viruses. Five active intein insertion sites (MCM-a through e) have been identified and studied in the archaeal replicative helicase gene *mcm*, making the MCM protein an ideal system for dissecting the dynamics of multi-intein genes. However, work in this system thus far has been limited to particular archaeal groups. To better understand the dynamics and diversity of these inteins, MCM homologs spanning all archaeal groups were extracted from NCBI's non-redundant protein sequence database, and the distribution and structural architectures of their inteins were thoroughly characterized. **Methods**: The amino acid sequences of 4,243 archaeal MCM homologs were retrieved from NCBI's non-redundant protein sequence database. These sequences were systematically assessed for their intein content through within-group multiple sequence alignments. To characterize the inteins present at each site, extensive intein structure predictions and comparisons were performed. Phylogenetic analyses were used to investigate intein relatedness between and within sites, as well as the distribution of different MCM inteins in geographically overlapping populations of archaea. **Results**: In total, 11 active MCM intein insertion sites were identified, expanding on the previously known five. The insertion sites have varied invasion activity levels across archaeal groups, with Nanobdellati (DPANN) being the only group with all 11 sites active. In all but two (Methanonatronarchaeia and Hadarchaeota) of the archaeal groups studied where inteins were present, there was at least one case of an MCM homolog invaded by more than one intein. With respect to intein structure, within-intein insertions bearing semblance to DNA-binding domains were identified, with varied presence between MCM inteins. Additionally, a study of archaeal MCM sequences of samples collected from the Atacama Desert in June 2013 revealed high MCM intein diversity levels. **Conclusions**: We present six new active intein insertion sites in archaeal MCM, expanding on the five previously known sites. No individual contained more than four MCM inteins simultaneously. Many inteins analyzed contained insertions bearing similarity to DNA-binding helix-turn-helix domains suggesting potential involvement in the intein homing process. Additionally, the high levels of MCM intein diversity observed in archaea from the Atacama Desert provide strong support for a co-existence model of intein persistence.

Keywords: inteins; multi-intein genes; archaea; replicative DNA helicase

1. Introduction

Inteins are mobile genetic elements invading highly conserved genes throughout all domains of life and viruses. An intein invades its host gene at the DNA level, similar to an intron [1]. Unlike introns which are spliced out at the RNA level, inteins splice themselves out at the protein level using an autocatalytic self-splicing reaction enabled by the intein's self-splicing domain. During protein splicing, the intein is able to seamlessly ligate the two halves of its host protein back together, allowing the host protein to function [2]. This natural capacity of inteins to engage in seamless protein

 doi:10.20944/preprints202505.2186.v1

splicing has made them invaluable tools in the development of protein engineering technology [3]. Recent large-scale intein characterization studies have revealed increasingly diverse intein architectures, such as varied architectures in the inteins of phages [4]. Such novel intein architectures hold the potential for new technological applications, emphasizing the importance of continued mass intein-characterization efforts.

Along with their novel biochemical capabilities, inteins also engage in unorthodox evolutionary behaviors. In addition to the self-splicing domain, full inteins contain a central homing endonuclease domain which bestows them with the ability to be inherited at super-Mendelian frequencies through the process of homing [5]. When in the presence of an uninvaded copy of the intein's host gene, the homing endonuclease domain will make a double-strand DNA break at the intein insertion site. Then, through homologous recombination-based DNA repair, the intein-containing copy of the gene can be used as a template leading the intein DNA sequence to be pasted into the previously uninvaded copy. As a result, homing allows inteins to rapidly proliferate through a population in spite of their fitness cost to the host [6]. Per the Goddard-Burt life cycle of a homing endonuclease driven selfish genetic element such as an intein, once the element reaches saturation in a population and there are no more empty target sites, the homing endonuclease is no longer under selective pressure to be maintained [5]. Once the homing endonuclease of an intein has severely decayed beyond function, it is referred to as a mini intein. In the Goddard-Burt model, the mini intein is eventually lost from the population. Recent models have expanded on the Goddard-Burt model to suggest the co-existence of the three states (intein-free, full intein-containing, and mini intein-containing) as opposed to a synchronized progression through the states [7,8], but to date no evidence of such co-existence of all three states in a single population has been shown [9].

There remains an incredible wealth of public sequence data to be explored for new intein architectures and evolutionary behaviors, particularly in archaea. Archaea contain inteins in a wide range of genes [10], but the majority of intensive archaeal intein characterization efforts have been focused on a few select groups such as haloarchaea. In addition to being a group ripe for the exploration of intein architectures and evolutionary dynamics, archaea offer a unique area of intein exploration in which both architecture and evolution can be studied: genes invaded by multiple inteins simultaneously [11,12]. The archaeal gene *mcm*, which encodes the MCM subunit of replicative DNA helicase [13,14], contains five intein insertion sites named MCM-a through e respectively. Inteins at sites MCM-a through d have been the subject of intein insertion site recognition and self-splicing investigations, particularly in haloarchaea [12,15]. Insertion site MCM-e is not invaded in any haloarchaea analyzed to date, but the site is active in some groups of non-haloarchaea. The MCM-e intein published to the intein database InBase 2.0 [16] in 2012 was from *Thermococcus litoralis*, with the insertion site name CDC21-e [17]. An analysis of the MCM inteins at sites MCM-a through d in haloarchaeal MCM homologs from the order Haloferacales revealed a wide array of intein invasion statuses (empty, single, double, triple, and quadruple), mini and full inteins at the same insertion sites in different homologs, and sporadic distribution of the four inteins across the host protein phylogeny [12]. The diversity of MCM inteins in this order alone begs the question of whether such patterns will hold when a similar analysis is performed on other archaeal lineages, and whether such diversity can also be found in a single group of geographically overlapping populations of archaea as opposed to a mass sampling of sequences from a wide array of timepoints and geographic locations.

To address these questions of intein architectural diversity, distribution patterns, and evolutionary dynamics at the population level, we gathered 4,243 complete archaeal MCM homologs from NCBI's non-redundant protein sequence database. To obtain as accurate a description of the MCM inteins across all archaea as possible with available data, an iterative search approach was used to thoroughly sample all available groups of archaea. A combination of sequence alignment and predicted structure-based analyses were used to characterize the inteins at all sites, through which six new archaeal MCM intein insertion sites were discovered. These new insertion sites all fall within the same catalytic ATPase domain of MCM as the known five (MCM-a through e). The sites are not

 doi:10.20944/preprints202505.2186.v1

active in all groups, with Nanobdellati (DPANN) being the only group to have at least one intein at all 11 MCM intein insertion sites. Our structural analyses revealed three sites within the MCM inteins where insertions resembling DNA-binding domains are found. These insertions vary in presence and size, adding a second facet to the inteins' architectural diversity beyond the status of their homing endonucleases (mini or full). Within this dataset were 26 haloarchaeal sequences from the Atacama Desert all sampled in June of 2013 from the same three locations as part of a metagenomic study by Finstad et al. [18]. This single group of geographically overlapping archaeal populations had greatly diverse MCM intein compositions, including no, mini, and full inteins at the same site in different individuals. Such a mixture of alleles strongly supports the co-existence model of intein persistence and captures the varied histories of inteins found at the different sites of a multi-intein gene.

2. Materials and Methods

Retrieving and curating amino acid sequence collection of archaeal MCM homologs. Using the MCM extein (host protein only, inteins removed in silico) sequence of *Haloferax mediterranei* (Protein Accession: WP_004058379.1) as the query sequence, PSI-BLAST [19] searches were performed against NCBI's non-redundant protein sequence database with maximum 500 target sequences and an e-value cutoff of 0.0001. No more than five iterations were allowed, and the resulting matches to be used for the subsequent iteration were manually pruned to remove partial MCM sequences (less than 600aa) and any non-MCM sequences. Each search was restricted to a different taxonomic group, such that effective sampling could be performed even for highly sequenced groups. After combining the smaller subsets of matches into taxonomically relevant groups (i.e., combining the four orders of Haloarchaea into a single Haloarchaea subgroup), 16 subgroups were formed: Haloarchaea (taxid 183963), Methanomicrobia (taxid 224756), Methanoliparia (taxid 2545688) Archaeoglobi (taxid 183980), Methanonatronarchaeia (taxid 171536), Thermoplasmatota (taxid 2283796), Nanohaloarchaea (taxid 1051663), Nanobdellati (DPANN) (taxid 1783276), Theionarchaea (taxid 1980645), Methanofastidiosa (taxid 1705400), Thermococci (taxid 183968), Hadarchaeota (taxid 3055124), Thermoproteati (TACK) (taxid 1783275), Promethearchaeati (Asgard) (taxid 1935183), and Hydrothermarchaeota (taxid 1935019).

Combined sequence and structure-based approach for characterizing the architectures of all inteins at each insertion site. For each of the 16 sets of MCM homologs, the sequences were initially aligned using MUSCLE [20] in SeaView [21] with slight manual adjustments to clarify intein versus extein (host protein) boundaries. For more complex cases such as Nanobdellati (DPANN) where all 11 intein insertion sites are active, and to varying degrees, no tried alignment algorithms (MUSCLE, clustalo [22], and MAFFT [23]) were able to properly align the sequences. For these cases, more extensive manual adjustments were required to establish the intein-host protein boundaries. These alignments were never directly used for phylogenetic reconstruction, and rather used to establish boundaries between host protein and intein sequences which were then extracted and re-aligned algorithmically for further analyses. For each intein insertion site within each taxonomic group sampled, the largest intein at the site was extracted, de-aligned, and used as input for AlphaFold3 [24]. Through this process, three sites within the inteins which occasionally contained insertions were identified: Insert Site 1 at the end of the N-terminal portion of the self-splicing domain, Insert Site 2 at the start of the C-terminal portion of the self-splicing domain, and Insert Site 3 ~12aa after Insert Site 2. Guided by the predicted structure of the largest intein, the homing endonuclease LAGLIDADG motif blocks and any within-intein insertions (Insertions 1, 2, and/or 3) were marked as selectable Sites in the sequence alignment in SeaView. By defining these Sites, each intein could be characterized based on its homing endonuclease and insertion architecture. The insertions were categorized as either small (20aa-60aa) or large (greater than 60aa) to capture the size variation observed between insertions at the same sites in different inteins. The minimum cutoff of 20aa was chosen based on the minimum length of a helix-turn-helix DNA-binding domain [25]. The sequence alignments for each group with declared Sites are provided as .mase files (viewable in SeaView) in **Supplemental Data 1.** The NCBI Protein Accession numbers are provided in the annotation line of every sequence. The

 doi:10.20944/preprints202505.2186.v1

inteins at each site were extracted into joined files, where Sites indicating Inserts 1, 2, and 3 were established in SeaView (.mase files available in **Supplemental Data 2**).

Unrooted amino acid sequence phylogenies. All phylogenies generated for this work should be considered unrooted, and are arbitrarily rooted when presented as such. For construction, the respective alignments were used as input for IQ-TREE2 [26], allowing ModelFinder [27] to identify the best fit model, and with 1000 replicates of ultrafast boostrapping [28]. The selected models are provided in the figure legends for each respective phylogeny. Treefiles were visualized using FigTree v.1.4.4 and Inkscape v.1.2.2.

3. Results

Analysis of MCM homologs across archaea reveals six new active MCM insertion sites. To investigate the abundance, structural features, and distribution of archaeal MCM inteins, 4,243 archaeal MCM homologs from NCBI's non-redundant protein sequence database were systematically collected. The domain Archaea was divided into subgroups following NCBI's Taxonomy Browser classifications, with more heavily sampled groups such as Stenosarchaea broken down into smaller groups for more thorough sampling. With thorough manual inspection of the sequence alignments generated for each subgroup, 11 distinct MCM intein insertion sites were identified (**Figure 1**). To our best knowledge, the only previously reported archaeal MCM intein insertion sites were MCM-a, b, c, d, and e. The new sites are all located in the same catalytic region as the known five (**Figure 1A-C**), owing to inteins' propensity to invade highly conserved regions [29]. The insertion sites cluster around especially important motifs for ATP binding by MCM subunits: the Walker A, Walker B, and arginine finger motifs [30]. Following the naming convention used for these intein insertion sites thus far, with slight alteration due to the very close proximity of two new sites to two pre-existing sites, we refer to these new sites as MCM-f, MCM-g, MCM-h, MCM-i, and MCM-d1 and MCM-e1. MCM-f through i are named in order of discovery and not their position in the linear sequence (**Figure 1D**), as this has been followed for naming the previously known sites. MCM-d1 and e1 are distinctly different from but very close (1 residue upstream) to MCM-d and e, thus we felt it beneficial to stray slightly from the traditional naming convention to reflect this. In this work, we refer to the original MCM-d and e as MCM-d2 and e2 due to them being one residue downstream of MCM-d1 and e1 respectively. Due to lack of sequence variation in the three MCM-d1 inteins, using phylogenetic reconstruction to further cement them as inteins of a separate insertion site than the MCM-d2 (d) inteins was not possible using an alignment of the MCM-d1 and d2 (d) inteins. However, there was sufficient variation among the inteins found at MCM-e1, allowing all MCM-e1 and MCM-e2 (e) inteins to be extracted, re-aligned, and used for phylogenetic reconstruction (**Figure S1**). In the resulting phylogeny, the MCM-e1 inteins group together as opposed to grouping with the MCM-e2 inteins from their respective archaeal groups (Nanobdellati (DPANN) and Promethearchaeati (Asgard)), providing further support for the MCM-e1 inteins being distinct from MCM-e2 (e).

 doi:10.20944/preprints202505.2186.v1

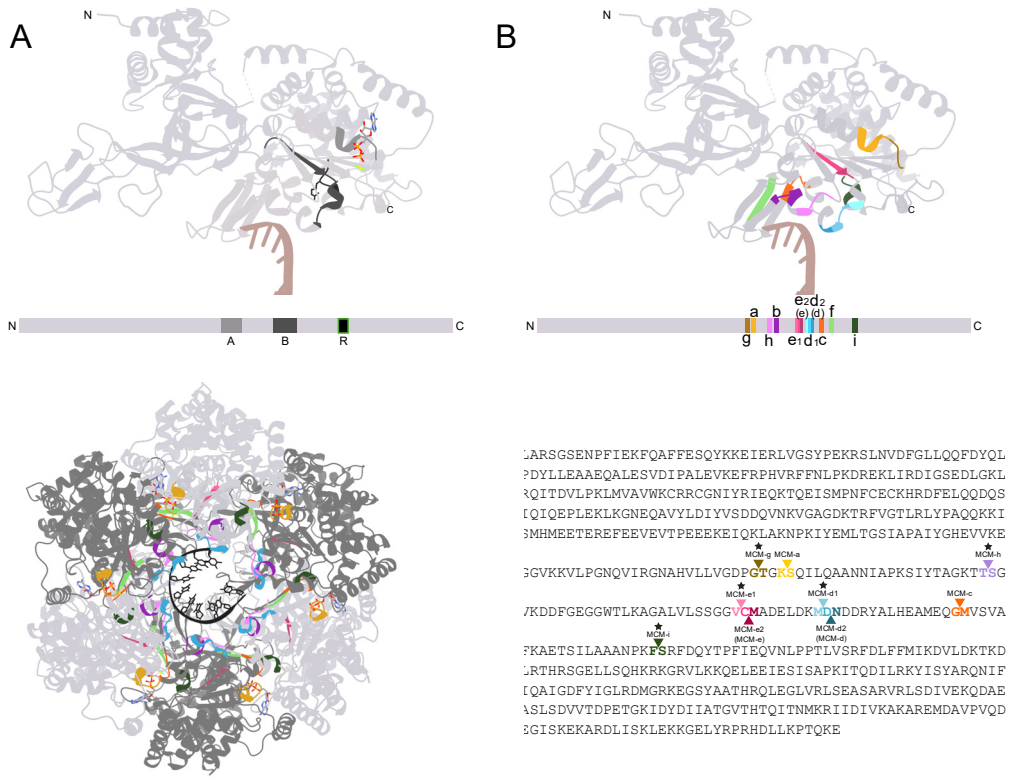
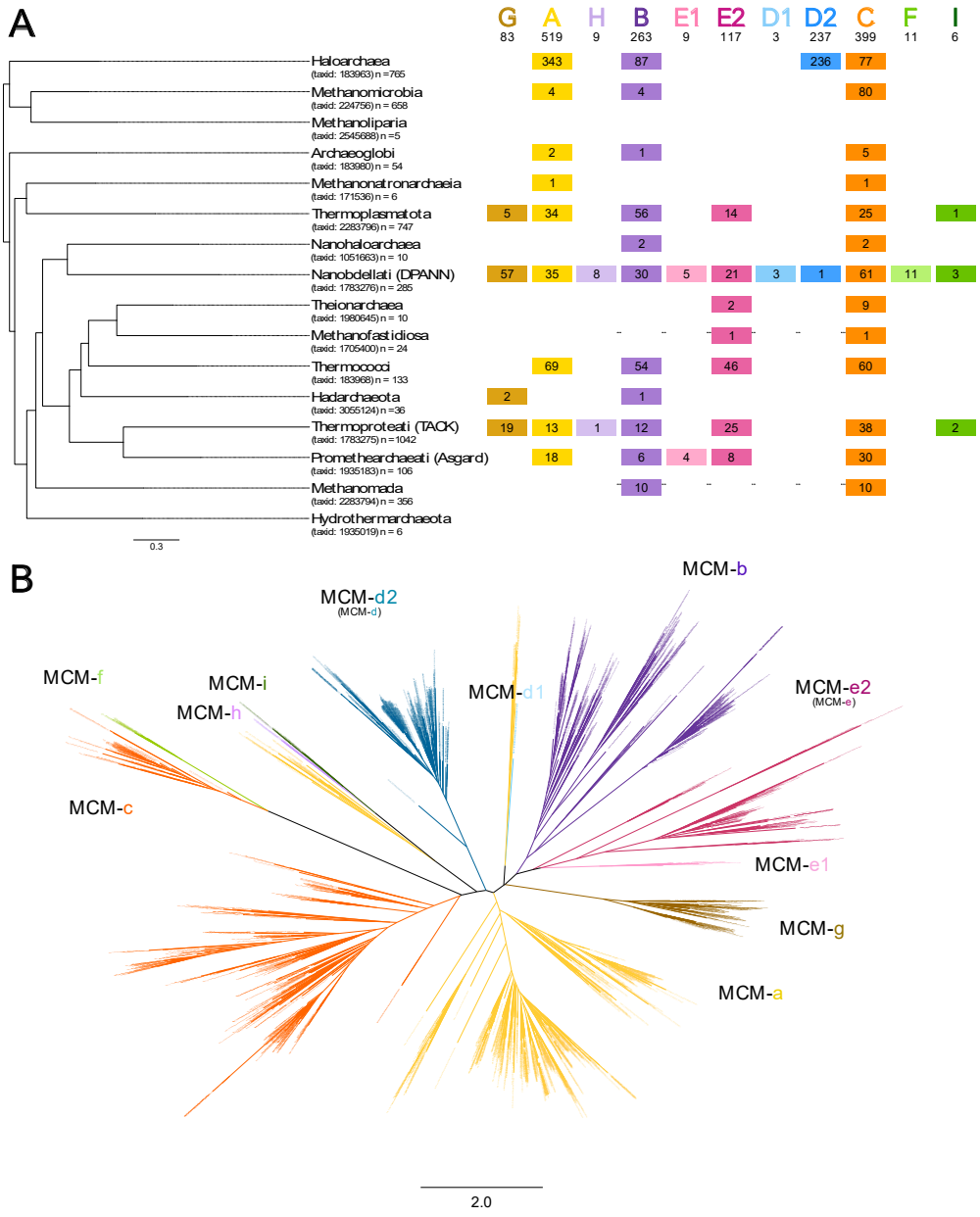


Figure 1. Six newly-discovered MCM intein insertion sites all fall within same catalytic ATPase domain as the five known insertion sites. A monomer from the solved crystal structure of the MCM homohexamer from archaeon *Saccharolobus solfataricus* P2 bound to ATP and single-stranded DNA (PDB 6MII Chain A [31]) was used to visualize core catalytic motifs (A) and all 11 intein insertion sites (B) of archaeal MCM. A. Walker A (A), Walker B (B), and arginine finger (R) motifs in structural and linear amino acid sequence contexts. Single stranded DNA (brown) and bound ATP molecules from the structure are visualized. The arginine side chain of the arginine finger is outlined in green. B. The five known (MCM-a, b, c, d, and e) and six new (MCM-d1, e1, f, g, h, and i) archaeal MCM intein insertion sites. Single stranded DNA (brown) from the structure is visualized. MCM-d and e are referred to as MCM-d2 and e2 respectively throughout this work. C. The homohexamer of PDB 6MII [31] with all 11 sites visualized. The sites all reside in the catalytic ATPase domain of MCM. The subunits are all identical, and only depicted in two shades of gray to better visualize the interfaces between subunits. The single stranded DNA fed through the center of the structure is depicted in black. D. The MCM host protein amino acid sequence of a Diapherotrites archaeon (Protein Accession: MBN2067331.1) with each MCM intein insertion site indicated with a triangle. The new sites presented in this work (MCM-d1, e1, f, g, h, and i) are indicated with stars.

Varied invasion activity levels and distinct evolutionary histories at each MCM intein insertion site. After establishing the positions of all MCM intein insertion sites, the insertion activity levels for each site across each archaeal group were assessed (**Figure 2A**). Out of the 16 subgroups, 14 had at least one active MCM intein insertion site. The only group in which all 11 sites are active, meaning at least one homolog from the group contains an intein at that site, is Nanobdellati (DPANN). Nanobdellati (DPANN) is also the only group with an active MCM intein insertion site which is inactive in all other groups (MCM-f). In contrast to MCM-f whose activity is seemingly limited to Nanobdellati (DPANN), in all intein-containing groups except for Hadarchaeota, at least one homolog had an intein at MCM-c. All new MCM intein insertion sites are less populated with inteins than the previously known sites. Similarly, instances of multi-intein invasions more frequently involved with previously known sites, with all quadruple invasions involving sites MCM-a, b, and c, with the fourth occupied site either being MCM-d2 (d) or e2 (e) (**Table S1**). In total, ~73.5% of homologs (3125) had no inteins, ~17% (709) had one intein, ~7% (305) had two inteins, ~2% (79)

 doi:10.20944/preprints202505.2186.v1

had three inteins, and ~0.5% (25) had four inteins (**Table 1**). While having 11 intein insertion sites and accounting for an intein status of empty, mini, or full at each site yields 177,147 theoretically possible MCM intein combinations, only 105 were observed. Out of those 105, 37 of the arrangements were observed in only one homolog each. All observed combinations of MCM inteins and their occurrences are available in **Table S1**. In addition to assessing intein invasion levels at each site, phylogenetic analysis was performed to assess grouping patterns of the MCM inteins (**Figure 2B**). Inteins at sites MCM-g, b, e1, d1, d2 (d), and f are monophyletic. The MCM-e2 (e) inteins all group together, with the MCM-e1 intein group emerging from them, adding further support to the differentiation between MCM-e1 and e2 (e) inteins despite their insertion sites being a single residue apart. This analysis also strengthens confidence in the differentiation between MCM-d1 and d2 (d) inteins, as the MCM-d1 inteins exhibit evolutionary distance from the MCM-d2 (d) inteins. The MCM-d1 inteins emerge from a group of MCM-a inteins. An additional small group of MCM-a inteins from which the MCM-h and i intein groups emerge is observed, and the majority of MCM-a inteins group together. The MCM-c inteins all group together, with the MCM-f inteins emerging from them. Over all, the inteins group strongly by insertion site as opposed to archaeal group.



 doi:10.20944/preprints202505.2186.v1

Figure 2. Distribution of MCM inteins across sampled archaeal groups and phylogeny of all MCM inteins analyzed. A. An unrooted phylogeny of the amino acid sequences of MCM exteins arbitrarily chosen from each sampled archaeal group (left) with the total inteins observed at each of the 11 MCM intein insertion sites mapped to the leaves (right). The NCBI Taxonomy ID (taxid) and total number of homologs collected for each group are listed in the phylogeny taxa labels. The total number of inteins observed at each MCM intein insertion site over all are listed in bold under the letter representing each site (G for MCM-g, A for MCM-a, etc.). **B.** All analyzed MCM inteins were extracted, joined into one file, and re-aligned using MAFFT [23]. The alignment was used as input for IQ-TREE2, allowing ModelFinder to choose the best fit model (Q.pfam+F+I+R10 chosen according to Bayesian Information Criterion), and performing 1000 replicates of ultrafast bootstrapping. The resulting tree was visualized using FigTree and Inkscape to illustrate the insertion sites at which each group of inteins on the phylogeny are found. A collapsed version of the phylogeny with bootstrap support values is provided in **Figure S2**.

Table 1. Degree of MCM intein invasion across homologs. The total numbers of MCM homologs with each degree of intein invasion are given. The degree of invasion ranged from empty (no inteins), to quadruple (four inteins).

MCM Intein Invasion Status	Total Homologs with Invasion Status
Empty	3125
Single	709
Double	305
Triple	79
Quadruple	25

Decaying versus full homing endonucleases and insertions within inteins at three distinct sites generate architectural diversity. For each MCM intein insertion site in each of the 16 groups of homologs, all inteins were categorized as either mini (no detectable homing endonuclease domain) or full (detectable homing endonuclease domain with both LAGLIDADG motifs). Mini inteins were only identified at sites MCM-a, b, c, d2, e1, and e2. For those sites, there were between 1.5 and 8 times more full inteins found than mini inteins (Figure S3). Using a combined sequence and predicted-structure based approach to define the domains of the inteins found at each site, inteins with additional domains beyond a homing endonuclease and self-splicing domain were identified. These inserted domains were identified in both full and mini inteins. Across all 1,656 inteins investigated, three distinct sub-insertion sites within the intein were identified: Insertion 1 located at the end of the N-terminal portion of the self-splicing domain; Insertion 2 at the beginning of the C-terminal portion of the self-splicing domain; Insertion 3 located ~12aa downstream of Insertion 2, placing it just after a conserved beta-strand in the C-terminal portion of the self-splicing domain [32,33] (Figure 3). Accounting for both the intein's homing endonuclease and sub-insertion status (mini or full intein; no, small, or large Insert 1; no, small, or large Insert 2; no, small, or large Insert 3) a total of 17 architectural variants were identified. The distribution of these architectural variants across the MCM intein insertion sites was assessed (Figure 4). Certain MCM intein insertion sites exhibited little variation in the architecture of their inteins, such as MCM-g which contained only full inteins with no insertions. This homogeneity is not due to limited distribution, as MCM-g inteins are present across several archaeal groups: Thermoplasmatota, Nanobdellati (DPANN), Hadarchaeota, and Thermoproteati (TACK) (Figure 2). In contrast, site MCM-b contained seven architectural variants. Insertion 3 was only identified in inteins located at site MCM-d2 (d).

 doi:10.20944/preprints202505.2186.v1

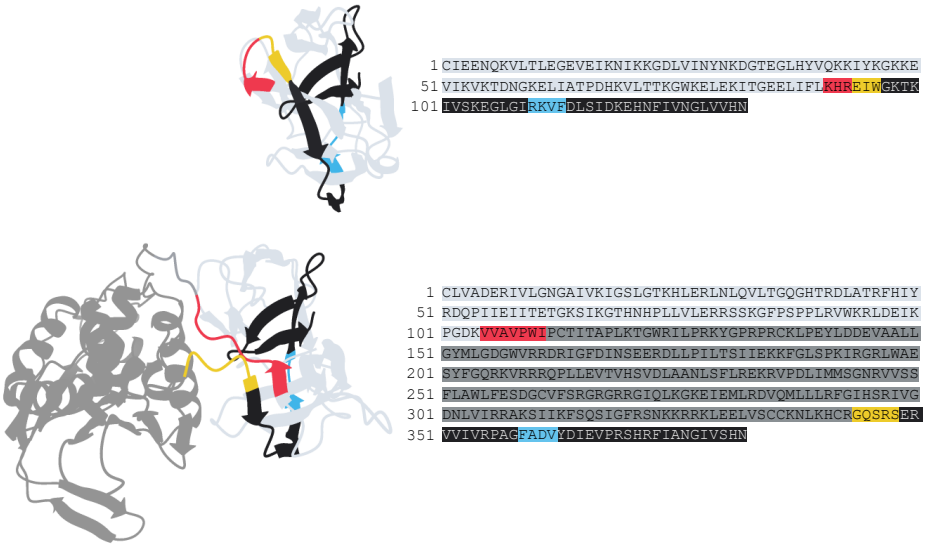


Figure 3. Three identified sub-insertion sites within the archaeal MCM inteins. AlphaFold3-predicted structures (left) and linear amino acid sequences (right) of the mini intein at insertion site MCM-e2 (e) from *Methanofastidiosum methylothiophilum* (Protein Accession KYC49087.1) (top) and the full intein at insertion site MCM-g from a Hadarchaeales archaeon (Protein Accession MEM2874594.1) (bottom). Neither of the inteins contain insertions. The N-terminal portion of the self-splicing domain is depicted in light gray, the homing endonuclease (only present in the full intein) in gray, and the C-terminal portion of the self-splicing domain in black. The relative positions of the sub-insertion sites identified are depicted in red, yellow, and blue for Insertions 1, 2, and 3 respectively. Sequence alignments clearly reveal differences between Insertion 1 and 2 in mini inteins (see **Supplemental Data 1**).

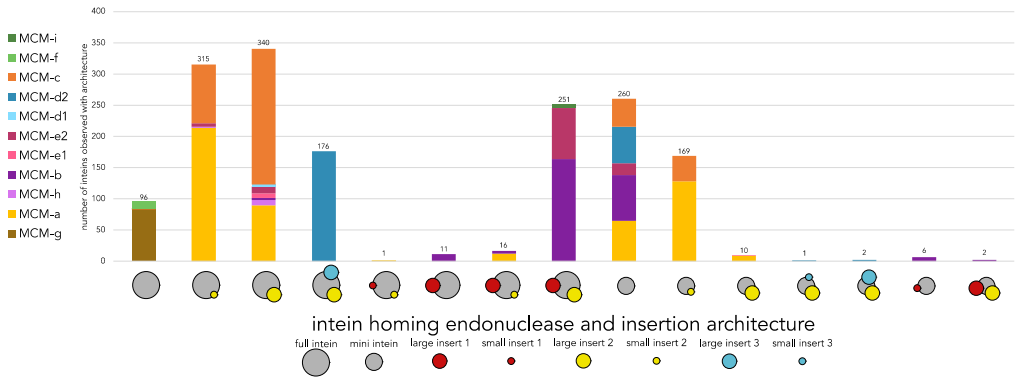


Figure 4. Distribution of observed intein architecture variants across the 11 MCM intein insertion sites. For each of the 17 observed combinations of homing endonuclease status (full or mini) and within-intein insertions (no, small, or large Inserts 1, 2, and/or 3), their distribution across the 11 MCM intein insertion sites was determined. The total number of inteins observed with a given architecture is included above the bar plot for that particular architecture.

Geographically overlapping populations in Atacama Desert have a wide range of MCM intein architectures and invasion statuses including co-existing empty, mini intein, and full intein alleles. To investigate models of intein persistence which involve co-existence of intein-free, full-intein, and mini-intein alleles [9], the haloarchaeal Atacama Desert sequences generated during the halite metagenome-based project of Finstad et al. [18] were utilized. All samples were collected from three regions in the Atacama Desert in Chile during June of 2013. From their sequence data, we identified 26 complete haloarchaeal MCM homologs. While the sequences are classified only as Halobacteriales

 doi:10.20944/preprints202505.2186.v1

archaea through the Finstad et al. project, we were able to provide more certainty on the genus-level identities of 24/26 archaea by comparing to sequences in our dataset of haloarchaea with known genus-level identities (Figure S4). By these classifications, these archaeal populations span the genera *Salinarchaeum, Natronomonas, Halovenus, Halostella, Halosimplex, Halosegnis, Halorussus, Halorubrum, Halomicrobium, Halomarina, Halococcus, Halobaculum*, and *Haloarcula*. Mapping the intein presence and architectures of these sequences onto a phylogeny of the MCM host proteins reveals a mixture of vertical inheritance and horizontal transfer, and varied intein architectures at a single site in closely related individuals (Figure 5). All degrees of MCM intein invasion except quadruple (empty, single, double, and triple) are observed in the population, as well as mini and full inteins at sites MCM-a and d. The Atacama Desert sequences provide concrete evidence for the co-existence of empty, mini-intein, and full-intein alleles in geographically overlapping populations of archaea from a single time period (June 2013), with each insertion site exhibiting different degrees of balance between the three alleles owing to their unique evolutionary histories.

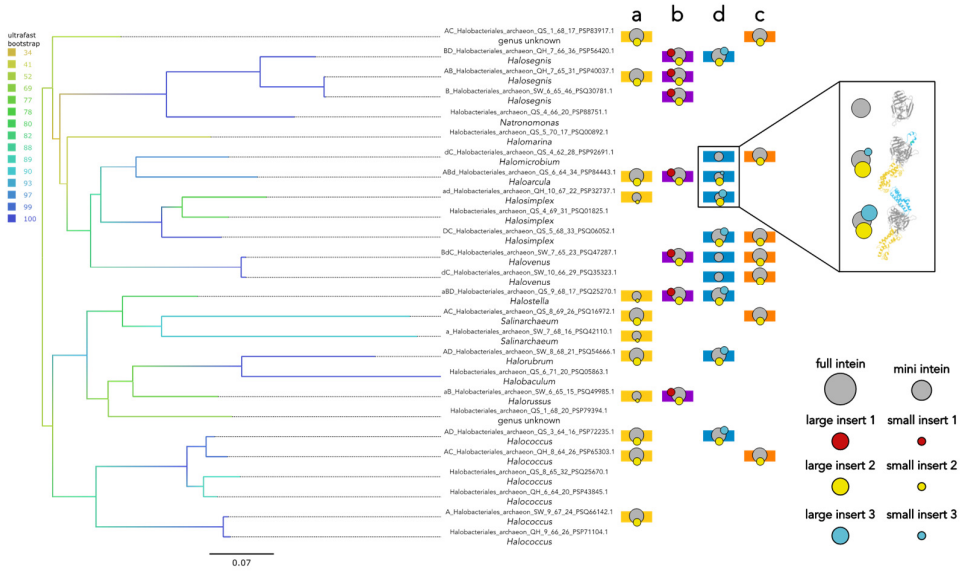


Figure 5. Distribution and architectures of MCM inteins across haloarchaea from Atacama Desert populations sampled by Finstad et al. The host protein sequences of the 26 haloarchaeal MCM homologs all from the Atacama Desert metagenomic project of Finstad et al. [18] in which all samples were collected in June of 2013 were used to construct an unrooted phylogeny. The host protein alignment was used as input with IQ-TREE2, allowing ModelFinder to choose the best fit model (LG+F+I+G4 chosen according to Bayesian Information Criterion), and performing 1000 replicates of ultrafast bootstrapping. The genera of the archaea were not provided by the original project, so comparisons to haloarchaeal MCM sequences of known genera were used to assign genera where confident assignments could be made (Figure S4). Circle diagrams are used to depict the intein architecture found at each MCM intein insertion site: a large central gray circle indicates a full intein with a homing endonuclease with both LAGLIDADG motifs intact; a small central gray circle indicates a mini intein with a degraded or completely lost homing endonuclease domain; large and small exterior circles indicate large (>60aa) and small (20aa-60aa) Insertions 1 (red), 2 (yellow), and 3 (blue). A clade with three particularly diverse mini inteins at MCM-d is used to demonstrate the associated AlphaFold3 predicted intein structures corresponding to the circle diagrams.

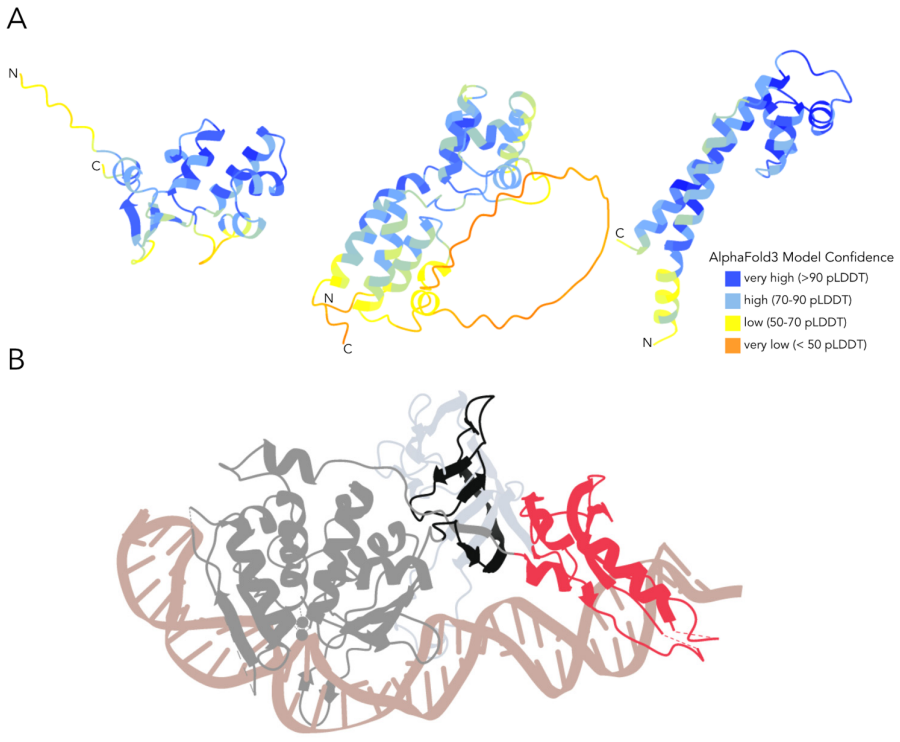
4. Discussion

Archaeal MCM is a powerful system for the continued exploration of multi-intein gene dynamics. Our work presents six previously unknown MCM intein insertion sites, and provides extensive characterization of the frequencies and architectures of inteins found at each site across

 doi:10.20944/preprints202505.2186.v1

archaea. We find the maximum degree for invasion of the *mcm* gene to be four inteins, despite many archaeal groups having more than four active MCM intein insertion sites. Similar caps on intein invasion have been observed for multi-intein genes such as the archaeal gene *polB* with up to three inteins simultaneously invading investigated copies of the gene [11] and a bacterial ribonucleotide reductase gene with up to four inteins [34]. Interestingly, the previous work in archaeal *polB* revealed *Haloquadratum walsbyi* to harbor the highest degree of intein invasion (triple), which is also the case for several strains of *Haloquadratum walsbyi* analyzed in this study (invaded at sites MCM-a, b, c, and d) making this a species of interest for future intein fitness cost investigations. However, the highest degree of invasion is still ultimately the rarest configuration in both *polB* and *mcm*. The additional sites presented in this work offer new avenues for exploring biochemical and molecular dynamics between inteins co-inhabiting a gene. In addition to the new insertion sites, the range of intein architectures discovered opens avenues for further investigation of the biochemical versatility of inteins with additional, potentially DNA-binding, domains.

Potential origin and role of the within-intein insertions. Archaea utilize several small DNA-binding proteins for transcriptional regulation, with some even interacting directly with MCM [35]. The core domain responsible for the DNA-binding abilities of these proteins is a helix-turn-helix domain [25], to which the sub-insertions found within many of the MCM inteins presented in this work bear strong resemblance in predicted structures (**Figure 6**). Thus, the pool of small DNA-binding helix-turn-helix proteins encoded in archaeal genomes could potentially be the source of some of the MCM intein sub-insertions. Within inteins, such additional domains have been observed in a region analogous to Insertion 1 in this work, at the end of the N-terminal portion of the self-splicing domain. In the crystal structure of the yeast vacuolar ATPase intein PI-*Sce*I (Protein Databank (PDB) entry 1LWS [36]), such an insertion is present. Work preceding the solving of PDB 1LWS implicated this region in DNA recognition and binding [37,38], with the crystal structure confirming direct interaction between this region and the DNA target sequence [36]. Thus, it is possible that the insertions within the MCM inteins are involved in the homing process, potentially in stabilizing the binding of the intein to its target DNA.



 doi:10.20944/preprints202505.2186.v1

Figure 6. Within-intein insertions with potential for DNA-binding capacity. A. The largest insertions found within the MCM inteins at Insert sites 1, 2, and 3 were used as input in AlphaFold3 (Inserts 1-3, left to right). The structures are colored by an AlphaFold-provided model confidence metric (predicted local distance difference test (pLDDT)), ranging from very high confidence (dark blue) to very low confidence (orange) [24]. **B.** The crystal structure of yeast vacuolar ATPase intein PI-*Sce*I bound to its DNA target (PDB 1LWS [36]). Following the scheme used in **Figure 3**, the N-terminal portion of the self-splicing domain is colored light gray, the DNA-binding insertion is colored red, the homing endonuclease is colored gray, and the C-terminal portion of the self-splicing domain is colored black. The DNA target is colored light brown.

Atacama Desert archaea provide support for co-existence model of intein persistence. Several models have been proposed to explain the life cycles of inteins in populations, with more recent proposals expanding on the Goddard-Burt homing cycle to suggest the co-existence of the three alleles (intein-free, full intein-containing, and mini intein-containing) in populations as a means of intein persistence [5,9]. Intein alleles of geographically overlapping populations at a single timepoint had not yet been assessed, and the Atacama Desert samples investigated in this work provide for the first time a clear picture of the MCM intein dynamics in a group of geographically overlapping archaeal populations. In these archaea, there is a balance of empty, mini, and full alleles for sites MCM-a and d, and empty and full alleles for sites MCM-b and c (the other seven sites are inactive, which is true of all haloarchaea analyzed). Thus, the MCM inteins in these populations operate in a manner more in line with a co-existence model for intein persistence [7–9], rather than the synchronized Goddard-Burt life cycle [5]. Future investigations of intein dynamics within geographically overlapping populations will continue to shed light on the frequencies of co-existence versus synchronized progression modes for intein persistence in natural populations.

5. Conclusions

Through this work, we characterized six new active intein insertion sites in the MCM subunit of archaeal replicative DNA helicase. By thoroughly characterizing the distributions and architectures of inteins found at all of the archaeal MCM intein insertion sites, we observed varying degrees of site activity between archaeal groups and a wide range of structural architectures. The insertions responsible for part of this architectural range bear similarity to DNA-binding proteins, suggesting a potential role in intein homing. Additionally, the MCM intein diversity observed in archaea from the Atacama Desert support a co-existence model of intein persistence in these populations, wherein each MCM intein insertion site exhibits a different arrangement of balance between empty, mini, and full alleles. Future endeavors to characterize the MCM intein content of geographically overlapping archaeal populations, as well as the intein content of other multi-intein genes, will continue to build our understanding of how intein distributions are balanced to maintain intein persistence.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1: Unrooted phylogeny of MCM-e1 and MCM-e2 (MCM-e) inteins; Figure S2: Collapsed MCM intein phylogeny with support values; Figure S3: Total mini and full inteins observed at each MCM intein insertion site for each archaeal group; Figure S4: Unrooted phylogeny of haloarchaeal MCM host proteins with Atacama Desert sequences indicated; Table S1: All observed MCM site combinations; Supplemental Data 1; Supplemental Data 2.

Author Contributions: Conceptualization, D.A. and J.P.G.; methodology, D.A., G.F.S., and J.P.G.; software, D.A., G.F.S., and J.P.G.; validation, D.A., G.F.S., and J.P.G.; formal analysis, D.A. and G.F.S.; investigation, D.A., G.F.S., and J.P.G.; resources, D.A., G.F.S., and J.P.G.; data curation, D.A. and G.F.S.; writing—original draft preparation, D.A., G.F.S., and J.P.G.; writing—review and editing, D.A., G.F.S., and J.P.G.; visualization, D.A.; supervision, D.A., G.F.S., and J.P.G.; project administration, J.P.G.; funding acquisition, J.P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

 doi:10.20944/preprints202505.2186.v1

Institutional Review Board Statement: Not applicable

Data Availability Statement: The protein sequences analyzed in this study are publicly available on NCBI's Protein Sequence Database. The individual Protein Accession numbers for each sequence are available in the annotations for each sequence in the provided alignments (**Supplemental Data 1**).

Acknowledgments: We thank the Computational Biology Core at the University of Connecticut for maintaining and managing the Xanadu computing clusters, which were a critical resource for many of the analyses completed in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

aa	amino acid
LAGLIDADG	one letter abbreviations of amino acids in a particular homing endonuclease motif
MCM	minichromosome maintenance
PDB	Protein Databank
pLDDT	predicted local distance difference test
taxid	Taxonomy ID

References

1. [1] R. Hirata, Y. Ohsumk, A. Nakano, H. Kawasaki, K. Suzuki, and Y. Anraku, "Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae.," *Journal of Biological Chemistry*, vol. 265, no. 12, pp. 6726–6733, Apr. 1990, doi: 10.1016/S0021-9258(19)39210-5.
2. [2] P. M. Kane, C. T. Yamashiro, D. F. Wolczyk, N. Neff, M. Goebl, and T. H. Stevens, "Protein Splicing Converts the Yeast TFP1 Gene Product to the 69-kdDSubunit of the Vacuolar H+-Adenosine Triphosphatase," *Science*, vol. 250, no. 4981, pp. 651–657, Nov. 1990, doi: 10.1126/science.2146742.
3. [3] H. Wang, L. Wang, B. Zhong, and Z. Dai, "Protein Splicing of Inteins: A Powerful Tool in Synthetic Biology," *Front. Bioeng. Biotechnol.*, vol. 10, p. 810180, Feb. 2022, doi: 10.3389/fbioe.2022.810180.
4. [4] S. P. Gosselin, D. Arsenault, and J. P. Gogarten, "Actinobacteriophage Inteins: Host Diversity, Local Dissemination, and Non-Canonical Architecture," Jan. 03, 2025, *bioRxiv*. doi: 10.1101/2025.01.02.630785.
5. [5] M. R. Goddard and A. Burt, "Recurrent invasion and extinction of a selfish gene," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 24, pp. 13880–13885, Nov. 1999, doi: 10.1073/pnas.96.24.13880.
6. [6] A. Naor et al., "Impact of a homing intein on recombination frequency and organismal fitness.," *Proc Natl Acad Sci U S A*, vol. 113, no. 32, pp. E4654-4661, Aug. 2016, doi: 10.1073/pnas.1606416113.
7. [7] A. Barzel, U. Obolski, J. P. Gogarten, M. Kupiec, and L. Hadany, "Home and away- the evolutionary dynamics of homing endonucleases," *BMC evolutionary biology*, vol. 11, p. 324, 2011, doi: 10.1186/1471-2148-11-324.
8. [8] K. Yahara, M. Fukuyo, A. Sasaki, and I. Kobayashi, "Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer," *Proc Natl Acad Sci U S A*, vol. 106, no. 44, pp. 18861–6, 2009.
9. [9] J. P. Gogarten and E. Hilario, "Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements," *BMC Evol Biol*, vol. 6, no. 1, p. 94, 2006, doi: 10.1186/1471-2148-6-94.
10. [10] O. Novikova et al., "Intein Clustering Suggests Functional Importance in Different Domains of Life," *Molecular Biology and Evolution*, vol. 33, no. 3, pp. 783–799, Mar. 2016, doi: 10.1093/molbev/msv271.
11. [11] A. Naor, R. Lazary, A. Barzel, R. T. Papke, and U. Gophna, "In Vivo Characterization of the Homing Endonuclease within the polB Gene in the Halophilic Archaeon Haloferax volcanii," *PLoS ONE*, vol. 6, no. 1, p. e15833, Jan. 2011, doi: 10.1371/journal.pone.0015833.

 doi:10.20944/preprints202505.2186.v1

12. [12] I. Turgeman-Grott et al., "Neighboring inteins interfere with one another's homing capacity," *PNAS Nexus*, vol. 2, no. 11, p. pgad354, Nov. 2023, doi: 10.1093/pnasnexus/pgad354.
13. [13] A. S. Brewster and X. S. Chen, "Insights into the MCM functional mechanism: lessons learned from the archaeal MCM complex," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 45, no. 3, pp. 243–256, Jun. 2010, doi: 10.3109/10409238.2010.484836.
14. [14] G. T. Maine, P. Sinha, and B.-K. Tye, "Mutants of S. cerevisiae defective in the maintenance of minichromosomes," *Genetics*, vol. 106, no. 3, pp. 365–385, Mar. 1984, doi: 10.1093/genetics/106.3.365.
15. [15] V. R. Yalala, A. K. Lynch, and K. V. Mills, "Conditional Alternative Protein Splicing Promoted by Inteins from Haloquadratum walsbyi," *Biochemistry*, vol. 61, no. 4, pp. 294–302, Feb. 2022, doi: 10.1021/acs.biochem.1c00788.
16. [16] "InBase2.0." Accessed: May 23, 2025. [Online]. Available: https://inbase.ligsciss.com/index.php?r=site/index
17. [17] F. B. Perler, "InBase: the Intein Database," *Nucleic Acids Research*, vol. 30, no. 1, pp. 383–384, Jan. 2002, doi: 10.1093/nar/30.1.383.
18. [18] K. M. Finstad et al., "Microbial Community Structure and the Persistence of Cyanobacterial Populations in Salt Crusts of the Hyperarid Atacama Desert from Genome-Resolved Metagenomics," *Front. Microbiol.*, vol. 8, p. 1435, Jul. 2017, doi: 10.3389/fmicb.2017.01435.
19. [19] S. F. Altschul et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.
20. [20] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, no. 1, p. 113, Aug. 2004, doi: 10.1186/1471-2105-5-113.
21. [21] M. Gouy, E. Tannier, N. Comte, and D. P. Parsons, "Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation," *Methods in molecular biology (Clifton, N.J.)*, vol. 2231, pp. 241–260, 2021, doi: 10.1007/978-1-0716-1036-7_15.
22. [22] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Science*, vol. 27, no. 1, pp. 135–145, Jan. 2018, doi: 10.1002/pro.3290.
23. [23] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.
24. [24] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
25. [25] M. Pellegrini-Calace, "Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2129–2140, Apr. 2005, doi: 10.1093/nar/gki349.
26. [26] B. Q. Minh et al., "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.
27. [27] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, "ModelFinder: fast model selection for accurate phylogenetic estimates," *Nature Methods*, vol. 14, no. 6, pp. 587–589, Jun. 2017, doi: 10.1038/nmeth.4285.
28. [28] D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: Improving the Ultrafast Bootstrap Approximation," *Molecular Biology and Evolution*, vol. 35, no. 2, pp. 518–522, Feb. 2018, doi: 10.1093/molbev/msx281.
29. [29] K. S. Swithers, A. G. Senejani, G. P. Fournier, and J. P. Gogarten, "Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements," *BMC Evolutionary Biology*, vol. 9, no. 1, p. 303, Dec. 2009, doi: 10.1186/1471-2148-9-303.
30. [30] A. S. Brewster et al., "Crystal structure of a near-full-length archaeal MCM: Functional insights for an AAA+ hexameric helicase," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 51, pp. 20191–20196, Dec. 2008, doi: 10.1073/pnas.0808037105.

 doi:10.20944/preprints202505.2186.v1

31. [31] M. Meagher, L. B. Epling, and E. J. Enemark, "DNA translocation mechanism of the MCM complex and implications for replication initiation," *Nat Commun*, vol. 10, no. 1, p. 3117, Jul. 2019, doi: 10.1038/s41467-019-11074-3.
32. [32] K. V. Mills, M. A. Johnson, and F. B. Perler, "Protein Splicing: How Inteins Escape from Precursor Proteins," *Journal of Biological Chemistry*, vol. 289, no. 21, pp. 14498–14505, May 2014, doi: 10.1074/jbc.R113.540310.
33. [33] K. Tori et al., "Splicing of the Mycobacteriophage Bethlehem DnaB Intein," *Journal of Biological Chemistry*, vol. 285, no. 4, pp. 2515–2526, Jan. 2010, doi: 10.1074/jbc.M109.069567.
34. [34] X.-Q. Liu, J. Yang, and Q. Meng, "Four Inteins and Three Group II Introns Encoded in a Bacterial Ribonucleotide Reductase Gene *," *Journal of Biological Chemistry*, vol. 278, no. 47, pp. 46826–46831, Nov. 2003, doi: 10.1074/jbc.M309575200.
35. [35] L. Aravind and E. V. Koonin, "DNA-binding proteins and evolution of transcription regulation in the archaea," *Nucleic Acids Research*, vol. 27, no. 23, pp. 4658–4670, Dec. 1999, doi: 10.1093/nar/27.23.4658.
36. [36] C. M. Moure, F. S. Gimble, and F. A. Quiocho, "Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence," *Nat Struct Biol*, vol. 9, no. 10, pp. 764–770, Oct. 2002, doi: 10.1038/nsb840.
37. [37] F. Christ, S. Steuer, H. Thole, W. Wende, A. Pingoud, and V. Pingoud, "A Model for the PI-SceI×DNA Complex Based on Multiple Base and Phosphate Backbone-specific Photocross-links," *Journal of Molecular Biology*, vol. 300, no. 4, pp. 841–849, Jul. 2000, doi: 10.1006/jmbi.2000.3872.
38. [38] D. Hu, M. Crist, X. Duan, F. A. Quiocho, and F. S. Gimble, "Probing the Structure of the PI-SceI-DNA Complex by Affinity Cleavage and Affinity Photocross-linking *," *Journal of Biological Chemistry*, vol. 275, no. 4, pp. 2705–2712, Jan. 2000, doi: 10.1074/jbc.275.4.2705.