Article

The evolutionary history of a DNA methylase reveals frequent horizontal transfer and within-gene recombination.

Sophia P. Gosselin [1], Danielle R. Arsenault [2], Catherine A. Jennings [3] and Johann Peter Gogarten [4,5,*]

1 Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA, Sophia.Gosselin@uconn.edu
2 Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA, Catherine.Jennings@uconn.edu
3 Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA, Danielle.Arsenault @uconn.edu
4 Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA,
5 Institute for Systems Genomics, University of Connecticut, Storrs, CT 06268-3125, USA; Gogarten@uconn.edu
* Correspondence: Gogarten@uconn.edu Tel.: 001 860 465 6267

Abstract: Inteins, often referred to as protein introns, are highly mobile genetic elements that invade conserved genes throughout the tree of life. Inteins have been found to invade a wide variety of key genes within actinophages. While in the process of conducting a survey of these inteins in actinophages we discovered that one protein family of methylases contained a putative intein, and two other unique insertion elements. These methylases are known to occur commonly in phages as orphan methylases (possibly as a form of resistance to restriction-modification systems). We found that the methylase family is not conserved within phage clusters and has a disparate distribution across divergent phage groups. We determined that two of the three insertion elements have a patchy distribution within the methylase protein family. We also found that the third insertion element is likely a second homing endonuclease, and that all three elements (the intein, the homing endonuclease, and what we refer to as the ShiLan domain) all have different insertion sites that are conserved in the methylase gene family. Furthermore, we find strong evidence that both the intein and ShiLan domain are partaking in long distance horizontal gene transfer events between divergent methylases in disparate phage hosts within the already dispersed methylase distribution. The reticulate evolutionary history of methylases and their insertion elements reveals high rates of gene transfer and within-gene recombination in actinophages.

Keywords: Actinophage; Inteins; LAGLIDADG Endonuclease; Homing; Horizontal Gene Transfer; Homologous Recombination; Selfish Genetic Elements.

1. Introduction

The SEA-PHAGES (Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science) Program organizes undergraduate courses in which students isolate, sequence, and annotate genomes of phages that infect actinobacteria. The SEA-PHAGES program is organized by Graham Hatfull's group at the University of Pittsburgh and the Howard Hughes Medical Institute's Science Education division [1]. As part of the bioinformatics section of the SEA-PHAGES course at the University of Connecticut, we developed a Course based Undergraduate Research Experience in which students characterized inteins and their evolution in actinophages.

Inteins, aka protein introns, are selfish genetic elements similar to self-splicing introns; the difference is that they are transcribed and translated together with the host protein, and only remove themselves following translation [2–6]. Typical inteins have two domains: the self-splicing domain acts to rejoin the two parts of the host protein, and the homing endonuclease domain allows the intein to invade previously uninvaded alleles.

 doi:10.20944/preprints202212.0517.v1

The homing endonuclease has a site specificity that corresponds to the intein insertion site and the surrounding nucleotides. It makes a double strand cut in the uninvaded allele at the site where the intein coding sequence is to be inserted. The hosts machinery then repairs the double strand break using the invaded allele as a template. Importantly inteins do not have their own mechanism to jump from one organism to another, rather they rely on the flow of genetic information that occurs by other mechanisms. Once two homologous genes, one with and one without an intein come together in one cell, the intein containing allele (if transcribed and translated) invades the intein free allele with high efficiency [7,8]. Intein alleles from different organisms that invade the same site in a gene are much more similar to one another than to inteins invading the same organism but at different sites or genes. Inteins invading the same site are known as intein alleles. Typically, inteins are found in conserved regions of conserved proteins [9]. This is also true for inteins in actinophages [10].

Actinophages are viruses that infect actinobacteria [11], a bacterial phylum that includees many soil bacteria (e.g., species in the genera *Streptomyces* and *Microbacterium*), but also important pathogens (e.g., *Mycobacterium tuberculosis* and *M. abscessus*). One driving force behind the study of actinophages is their potential use in phage therapy [12,13]. The discovery of phages that can lyse bacterial cultures let d'Herelle, one of the co-discoverers of bacteriophage, to use them to combat bacterial infections [14]. The ability of phages to effectively attack Gram positive bacteria is one of the driving forces behind the SEA-PHAGES project. On December 12th, 2022, the phagesdb databank [15,16] reported on 22387 phages of which 4184 have complete genome sequence records which are in turn divided into 148 clusters of related phages and 62 singletons, i.e., phages that currently are not members of a cluster. The proteins encoded in these genomes are grouped into families (called phams or phamilies) based on sequence similarity [17], but the membership of individual protein coding genes in phams changes as more sequences are added to the database. In general, current and past research find that these phamilies show typical patterns of intein invasion [9]. Inteins are found in important genes such as helicases, terminases, and many other proteins essential for capsid structure, DNA replication and packaging [10].

One exception to the preference for conserved and important proteins that we found was an intein in a gene annotated as a putative DNA methylase. In addition to the intein, a few members of the analyzed methylase family contain another region encoding a nearly identical protein sequence fragment, which in the following we label as the ShiLan domain, after the phage in which this insertion was first noticed. This region was present in otherwise divergent members of the gene family. A third sequence present in only a few members of the methylase family encoded an additional endonuclease of the LAGLIDADG family. This endonuclease was different from the LAGLIDADG homing endonuclease associated with the intein and showed significant similarity to a homing endonuclease in a group I intron.

DNA methylases play diverse roles. They often are part of Restriction Modification Systems (RMSs), and they play a role in marking DNA regions and are critical for mismatch repair and regulation of the origin of replication in bacteria [18,19]. RMSs are often considered part of a bacterial defense system, recognizing and restricting DNA with a different methylation pattern. However, RMSs are also a form of addiction cassette, encoding a toxin anti toxin system [20,21]. The toxin in this case is the restriction enzyme and the antitoxin is the methylase. If the associated methylase is lost from a cell's RMS, the remaining restriction enzyme activity will destroy the organism's genome. For an RMS to be lost, first the restriction activity needs to decay; only then can the whole system be deleted. In line with their characterization as addiction cassettes, RMSs are frequently encoded on plasmids and often have a disjunct distribution (e.g., [22,23]). RMSs come in 4 different varieties [24]. Type I RMSs are composed of three different polypeptides acting as a single complex. One peptide acts as a specificity recognition protein, and the other two act to modify or cleave the bound DNA. Type II RMSs are the simplest true RMSs, consisting of two separate proteins (one endonuclease, and one methytransferase) that can

 doi:10.20944/preprints202212.0517.v1

act independently of each other and of a specificity protein. Type III RMSs form a complex like Type I, but lack the associated specificity protein. Lastly there are Type IV RMSs which lack a modification protein entirely, and therefore do not count as true RMSs. However, there are some exceptions to this schema, where multiple activities are encoded on a single peptide (e.g., Type IIB RMSs [25]).

Our analysis of methylase sequences reveals a sporadic distribution of the methylases, frequent transfer of genes between phages belonging to different clusters, a surprising number of recombination events between the methylase sequences from divergent phages, and a recent intein invasion of phages isolated from the same geographical area.

2. Materials and Methods

Initial sequence discovery occurred during visual inspection of viral genomes via Phamerator [26] and via repeated searches for inteins in the PhagesDB database using psi-BLAST [27]. All sequences used in this research were downloaded from PhagesDB on 6/17/2022; metadata on the phages were updated on 8/31/2022. In-house scripts used to construct local databases and extract metadata can be found at https://github.com/sophiagosselin/Methylase_Insertions.

MAFFT (v7.471) [21] was used to create sequence alignments. MAFFT alignments were conducted using the globalpair and reorder settings, a maximum iteration count of 1000 for the compact alignment; and using an unalignlevel setting of 0.8. for the gappy alignment.

SEAVIEW (v5.0.4) [29] was used to inspect alignments and define insertion sequences.

Phylogenies were built using IQ-TREE (v2.1.3) [30]. As the different components of the methylase sequences (extein, intein, ShiLan domain, endonuclease domain) likely had different evolutionary histories, we estimated the best fitting model for each component separately using IQ-TREE's built in ModelFinder. Table 1 lists the models selected for each dataset. Bootstrap support was created using the ultrafast bootstrapping option with 1000 bootstraps.

Constrains for the AU-test, i.e., unresolved trees that represented the constrains, were created in a text editor, and the best maximum likelihood (ml) tree given these constrains was then constructed in IQ-TREE [30] using the –g option and constrained Newick trees. Constrained maximum likelihood trees were built such that the clan of interest was constrained to only contain members of this group, but all other taxa could be freely placed.

Resulting phylogenies were visualized in Figtree (v1.4.4) and then editorialized in vector graphics software (InkScape (v2.2)). Maximum likelihood distances between sequences were calculated with IQ-TREE and the correlation between pairwise distance matrices was calculated in Microsoft Excel (v16.67). These ml trees were then used in the AU-test.

A predicted protein structure was generated for the methylase from the PopTart phage. The PopTart methylase sequence was used as input for the AlphaFold v2.2.4. [31] Jupyter notebook hosted on Google Colab. The predicted structure was then colored in Chimera [32]to indicate the insertion sites of the ShiLan domain, intein, and second homing endonuclease. In addition, AlphaFold v2.2.4 was used to generate a predicted structure for the full methylase from the Taj phage. The Taj methylase does not contain the ShiLan domain or intein, but does contain the second homing endonuclease. The predicted structure was colored in Chimera to indicate the three insertion sites and the homing endonuclease domain.

HHPred [33] analyses were performed using the default settings of the webserver at https://toolkit.tuebingen.mpg.de. Databases searched were PDB_mmCIF70_12_Aug, Pfam-A_v35, NCBI_conserved_Domains(CD)_v3.19, and TIGRFAMs_v15.0.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

Table 1. Models selected by IQ-TREE using the Bayesian Information Criterion (BIC) for the different datasets

Alignment	**Best fitting model[$$]**
Exteins[$] compact alignment	VT+F+R7
Exteins[$] gappy alignment	WAG+R5
Inteins (compact alignment)	WAG+F+I
Exteins[$] (intein containing, compact alignment)	Blosum62+F+G4
ShiLan domain (compact alignment)	HIVb+F+I
Exteins (ShiLan domain containing, compact alignment)	WAG+G4
Second endonuclease domain	Q.pfam
Exteins (Second endonuclease domain containing, compact alignment)	WAG+G4

[$] Extein sequences excluded the intein, ShiLan, and the second endonuclease domains.
[$$] See the IQ-TREE manual [30] for detailed descriptions of the models.

3. Results

3.1. Distribution of homologs to the Dorotheee_75 methylase

The assignment of putative DNA methylases to *certain* phamilies in phagesdb changed over the course of the project. On May 18th 2022 the methylase encoded in the genome of phage Taj (Taj_79) was part of pham 105558 with 254 members. The intein-containing methylase from phage Dorothy (Dorothy_75)) was part of pham 106461 with 17 members. The latter pham contained most but not all of the intein containing homologs; however, some intein containing methylases had been placed into the former. The two phams undoubtedly contain homologous sequences. A pairwise comparison in PRSS [34] between the methylase from ShiLan (ShiLan_65 placed in the former cluster) with Dorothy_75 resulted in a 938 aa overlap with a Z-score of 3833 and an e-value (E(10000)) of $7*10^{-169}$. (E(10000) gives the expectation of the number of matches with the same or better quality if 10,000 shuffled sequences are compared). For the following analyses the two phams were combined. The homologous methylases are found in many different phage clusters, but in none of the clusters do all members contain a homolog Fig 1.

The homologous methylases have a wide distribution among actinophages. They are found in fourteen different phage clusters (Fig. 1). However, these methylases have a very sparse distribution. In none of the clusters is the methylase present in all members of the cluster, and in most clusters the phages without the methylase gene outnumber the ones that encode the methylase in their genome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

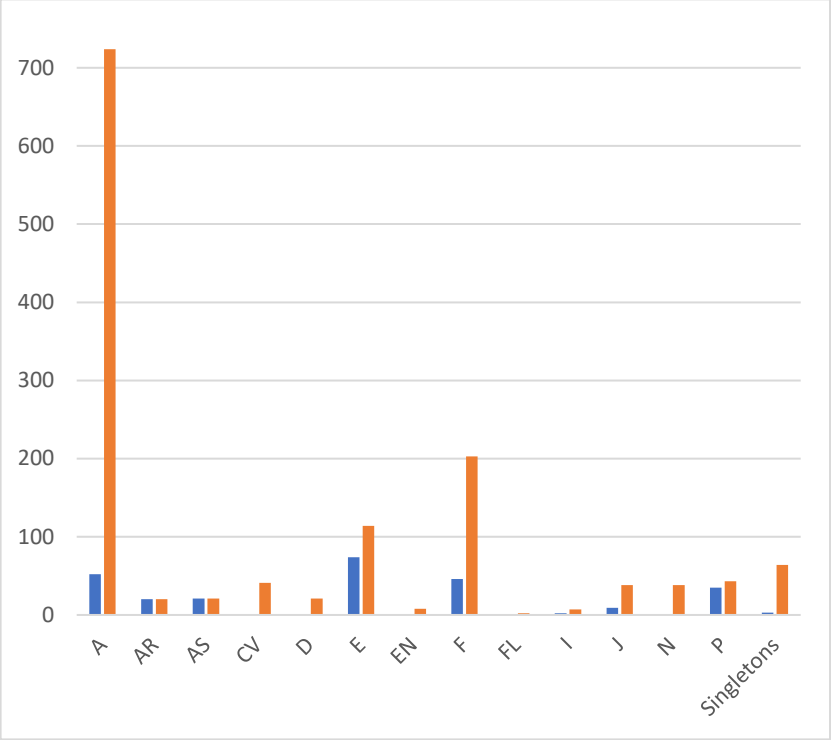

Figure 1. Bar graph giving the number of genomes per cluster (on Aug. 30th 2022) (orange), and the number of genomes that encode a homologous methylase (blue).

3.2. Alignments and insertion sequences

For our analyses we used two different alignments. One is a more compact traditional multiple sequence alignment (MSA), and the other is an alignment that aligns uncertain alignment regions to gaps in the other sequences. In the following we label these as the compact and gappy alignments respectively. The gappy alignment focusses on reliably aligned residues, minimizes potential artifacts from the guide tree and results in phylogenies with much shorter branch lengths since gaps in the other sequences are encoded as missing data.

Some of the methylases had been identified as intein containing using a PSI BLAST search. The multiple sequence alignment of the methylase sequences revealed that 21 of them contain an intein in the same position. The following findings confirm the identity of this insertion as an intein:

- The intein is present in only a fraction of the sequences.
- It is inserted in a conserved region of the methylases (Fig2A).
- Results from an HHPred search show homology to inteins over the whole length of the insertion(Fig 3).
- The intein sequences have a phylogeny different from the remainder of the methylase (see below).

In addition, we found another insertion in ShiLan_65 and seven other methylases. Going forward we will refer to this insertion as the ShiLan domain. The sequence of this insertion is conserved, is inserted in a conserved region of the methylase (Fig. 2B) and is found in divergent methylases of phages from the E and F clusters. An HHPred search using this sequence as a query resulted in only one low quality match to the beginning of the insertion sequence (Fig.3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

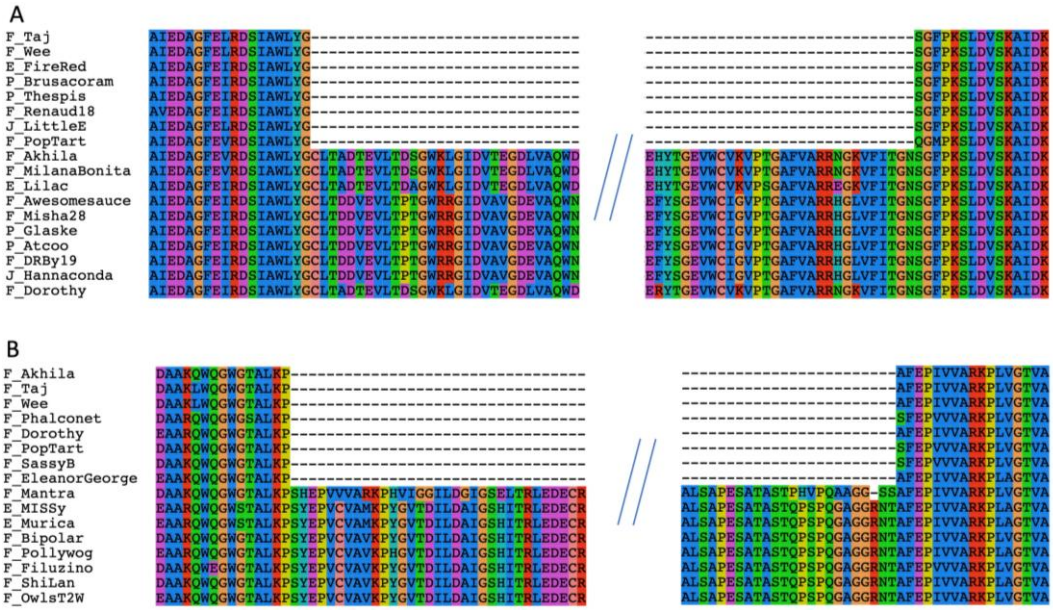

Figure 2. Alignment of the beginning and end of the Intein (panelA) and the ShiLan-Domain (panel B) with their associated surrounding regions. In phage Dorothy the Intein is 327 amino acids long and the ShiLan-domain in phage ShiLan is 202 amino acids long.

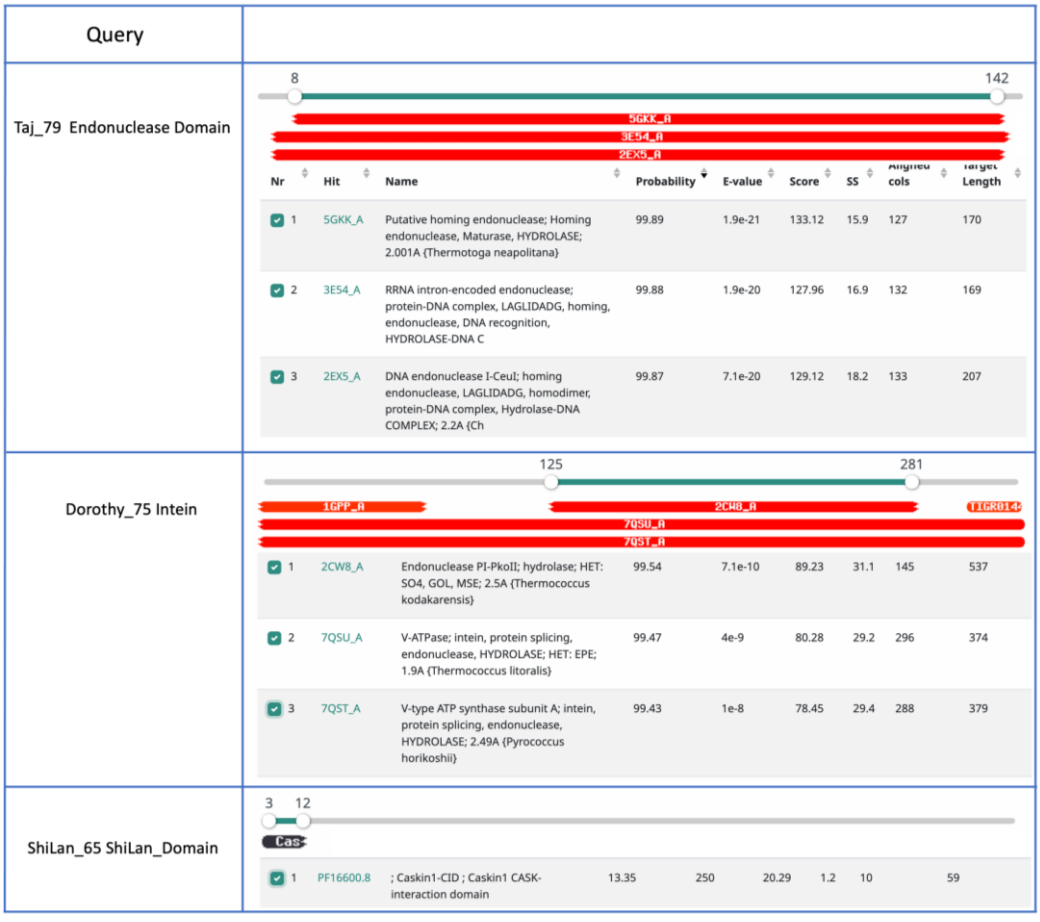

Figure 3: Results from HHPred searches using the three inserted sequences as queries. Only the three most probably matches are given. The endonuclease domain shows significant similarity to endonuclease of the LAGLIDADG type, the intein over its whole lengths matches intein sequences (match two and three). The search with the ShiLan domain did not result in any significant matches.

In an unrelated project, one of us (DRA) searched the NCBI Virus database using the LAGLIDADG homing endonuclease contained in a group I intron (accession #

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

YP_005089794) as the query. This intron is located in in the atpA gene of a *Dunaliella salina* chloroplast genome (accession # NC_016732). This homing endonuclease had a significant BLAST search hit to the Taj_79 methylase. Inspection of the match and the MSA of the methylases revealed that this endonuclease domain was well conserved in 9 methylase sequences. Each of these methylases contained the typical motif of a LAGLIDADG endonuclease, but surprisingly was located outside and upstream of the intein insertion site. In an HHPred search this endonuclease had significant matches to homing endonucleases (Fig. 3).

Fig. 4 gives the location of each insertion on an AlphaFold predicted structure of the methylases from phages PopTart, which does not contain any of the three insertion elements, and Taj, which contains the endonuclease insertion.

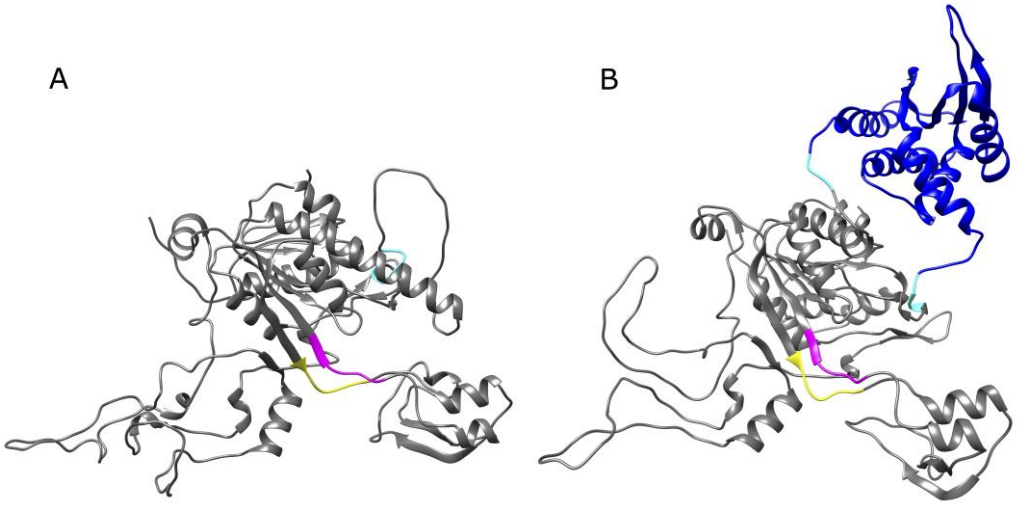

Figure 4: Structure of the PopTart_63 (panel A) and Taj_79 (panel B) methylase as predicted by AlphaFold v2.2.4. Three residues upstream and downstream of each element's insertion site are indicated as follows: ShiLan domain in magenta, intein in yellow, and the second LAGLIDADG homing endonuclease in cyan. The putative homing endonuclease domain in Taj_79 is in blue. The model confidence for the two structures is depicted in Fig. S2.

3.3. Methylase Phylogeny.

3.3.1. Methylases do not group according to the cluster in which the phages belong.

The maximum likelihood phylogenies (Fig. 5 and supplementary Figure 1A and B) reconstructed from the compact and gappy alignments are similar in that:

- The sequences from phage clusters DC, CV, FL, AR and AS together with two sequences from separate singletons form a well-supported clan in both phylogenies (a clan is a group of tips that group together in an unrooted phylogeny, corresponding to a clade in a rooted phylogeny).
- The sequences from clusters F, P, E, A and J do not form clans.

The two phylogenies differ in that:

- Details of the branching order are not consistent between the two topologies.
- The two phylogenies have different estimated branch lengths.

We used KH [35] , SH [36], and AU [37]-tests, as implemented in IQ-TREE2 [38] to determine if the possibility of the sequences from each cluster grouping together could be rejected with confidence. Bias created through the alignment process tends to reinforce the clusters from the guide tree. To minimize the effect of alignment bias, we chose the gappy alignment for this analysis. The maximum likelihood phylogeny constrained to group all the clusters as individual clans was rejected, as was the maximum likelihood phylogeny that only constrained methylase from cluster F as a clan. Results are summarized in Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

Table 2. Results from statistical test comparing constrained maximum likelihood trees to the overall best maximum likelihood tree determined by IQ-TREE [38]. Numbers give the probability with which the tree can be considered as part of the 95% confidence set. Trees rejected as being part of the 95% confidence set are indicated in bold.

Tree	**p-KH** [35]	**p-SH** [36]	**p-AU** [37]
ml tree	0.319	0.84	0.372
all clusters constrained	**0.0006**	**0.001**	**5.41 E-05**
cluster A constrained	0.0947	0.325	0.0529
cluster AS constrained	0.46	0.928	0.612
cluster AR constrained	0.363	0.939	0.454
cluster E constrained	0.54	1	0.618
cluster F constrained	**0.0013**	**0.0027**	**0.00012**
cluster E constrained	0.23	0.609	0.171
cluster J constrained	0.299	0.81	0.335
cluster I constrained	0.418	0.921	0.484
all Intein containing seq.	**0**	**0**	**8.05 E-54**
all ShiLan Domain containing seq.	**0**	**0**	**3.11 E-06**
all Endonucl. Domain containing seq.	0.501	1	0.549

3.3.2. Intein and ShiLan domain containing methylases do not form clans.

Figure 5 depicts the maximum likelihood phylogeny calculated from the gappy alignment of the methylase family. The sequences containing the intein, the ShiLan domain and the additional endonuclease domains are highlighted. The intein and ShiLan domain containing sequences do not cluster together, whereas the sequences with the additional endonuclease domain are restricted to the F-cluster and group together as a clan.

We calculated the best maximum likelihood trees which constrained each of the three types of insertion sequences to its own clan. The trees constraining the intein or ShiLan domain containing sequences into a clan were confidently rejected as being part of the confidence set (Table 2). In contrast, the phylogeny constraining the sequences into a clan that harbors the additional LAGLIDADG endonuclease domain was not rejected.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

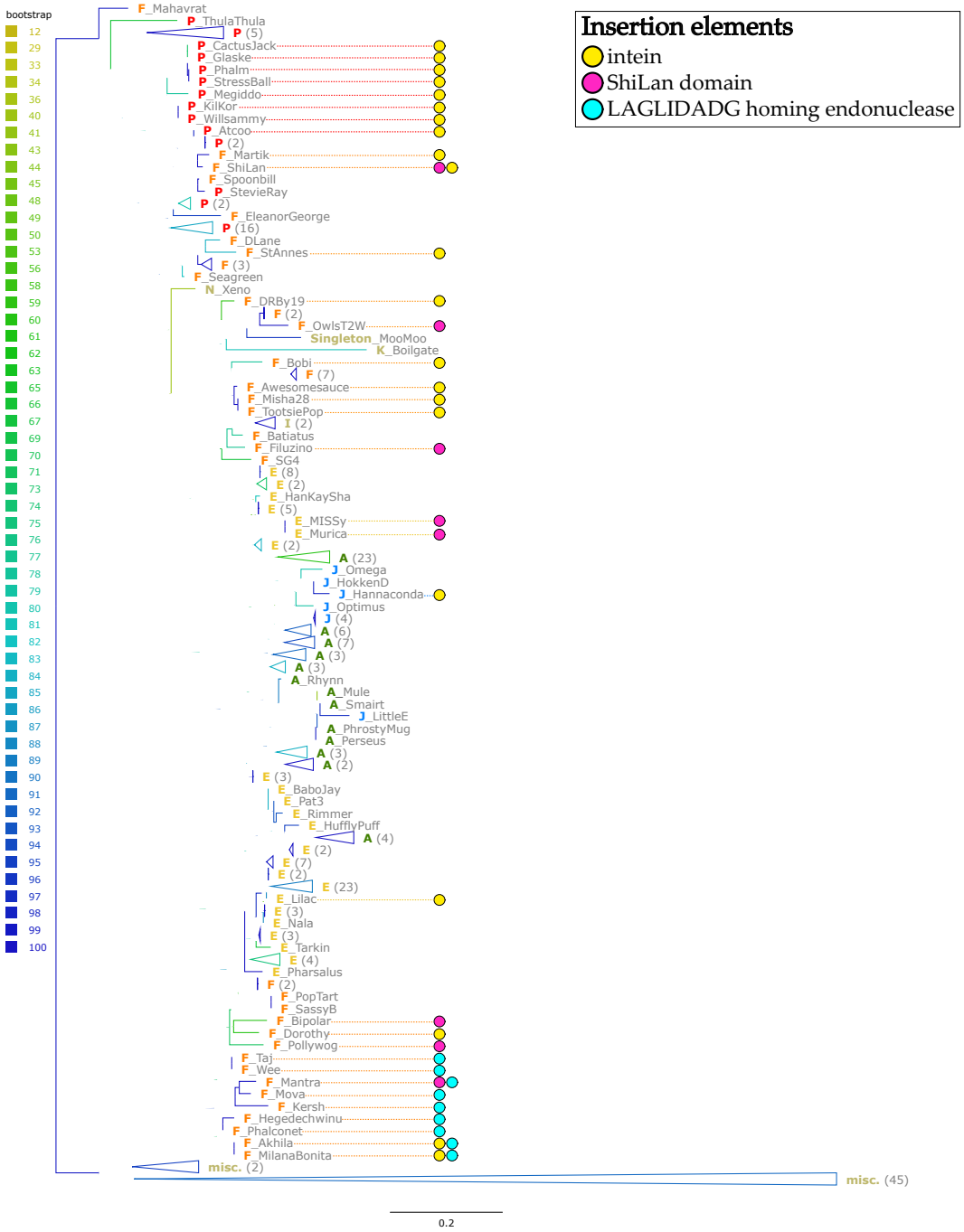

Figure 5: Phylogeny of the methylase family calculated with IQ-TREE from the gappy alignment with the insertion sequences (intein, ShiLan, and second endonuclease domain) removed. The cluster to which each phage belongs is denoted by a colored prefix in bold. Sequences that contain the intein (yellow), the ShiLan domain (magenta) and the second homing endonuclease domain (cyan) are indicated by colored circles. Branches are colored to reflect bootstrap support values. Clades of insertion-free sequences belonging to the same phage cluster were collapsed where possible to increase readability. Numbers in parentheses indicate the number of taxa contained within the collapsed clade. Collapsed clades labeled "misc." contain insertion-free taxa from several phage clusters (AR, AS, CV, DC, EN) and three singletons. Supplementary Figure S1B contains the phylogeny without collapsed branches.

3.3.3. Comparison of phylogenies for the inserted elements and the methylases that harbor these elements.

Phylogenies for the intein, the ShiLan and the second endonuclease domains were compared to the extein sequences (minus the intein, second endonuclease, and ShiLan domains) that harbored the respective elements (Figure 6). To better capture the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2022 doi:10.20944/preprints202212.0517.v1

divergence of the sequences, the constrained alignment was used for these comparisons. The intein sequences fall into two well supported groups (colored blue and red in Fig.6 B); these two groups do not form clans in the extein phylogeny (Fig. 6A). The divergence within the two intein groups is also much less than the divergence between the extein sequences (in Fig. 6A names are colored according to the two intein groups).

For both the second endonuclease (Fig. 6 E) and the ShiLan domain (Fig. 6 H) one of the sequences (from phages Kersh and Mantra, respectively) is more divergent, while the remaining sequences are much more similar to one another. These more related sequences are less divergent from one another than the methylase sequences in which they are found. For all three elements, the distances between the elements do not correlate with the distances between the methylases that contain said elements (Fig 6 C, F, and I). The R squared values are .051, .086, and 0.001 for the correlation between intein and extein, endonuclease and methylase, and ShiLan domain and methylase, respectively. However, when the distances to the most divergent endonuclease domain (from phage Kersh) is excluded, the R squared for the second endonuclease increases to 0.69, whereas the R square for the ShiLan domain with the distances to phage Mantra omitted remains low at 0.12.

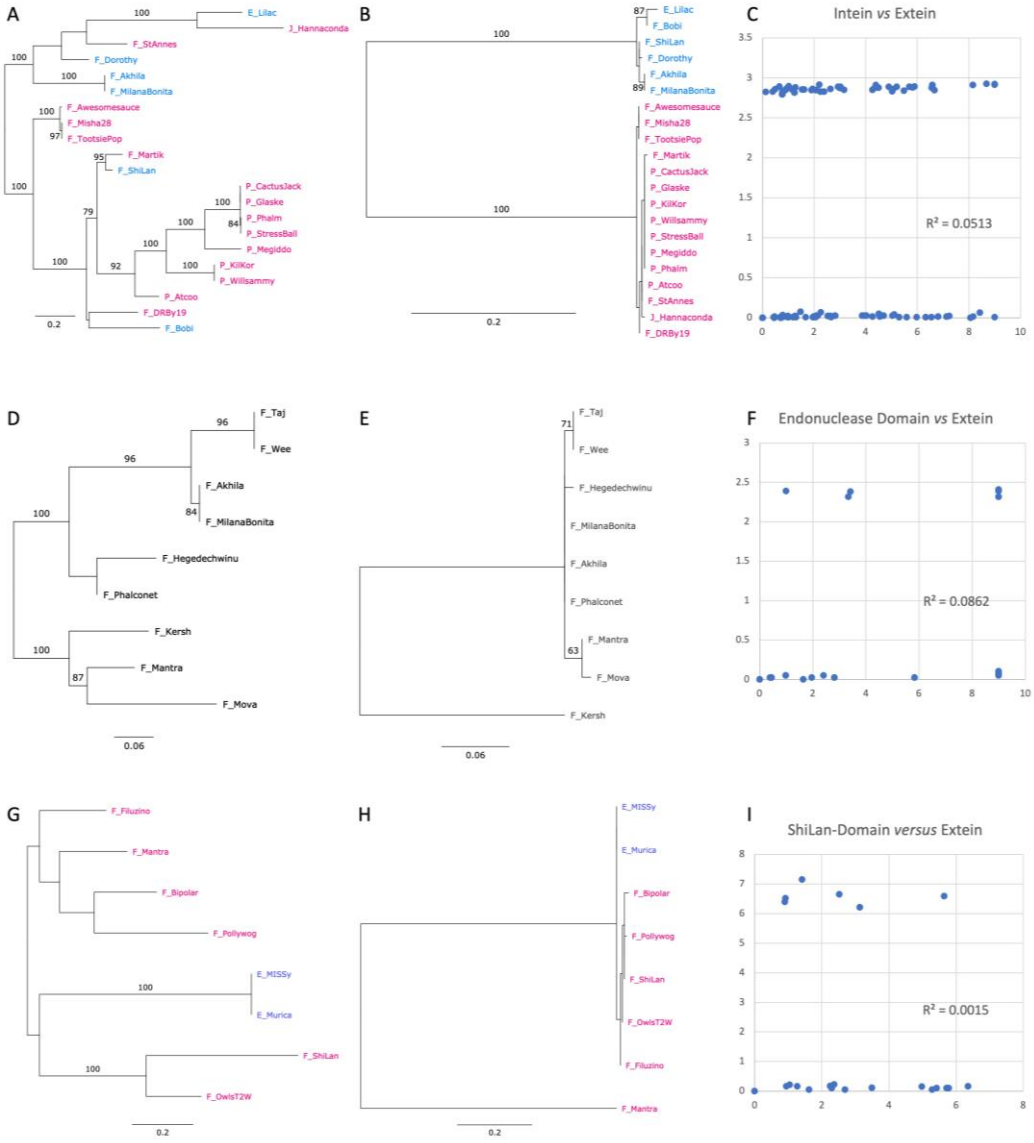

Figure 6: Phylogeny of subsets of the methylase family (panel A, D, G) compared to the phylogenies of the intein (B), second endonuclease (E) and ShiLan domain (panel H). The phage names in panels (A) and (B) are colored to reflect the two intein groups. Phage names in panel (G) and (H) are colored

 doi:10.20944/preprints202212.0517.v1

to reflect the clusters to which the phage belongs. Panels (C), (F), and (I) plot the pairwise maximum likelihood (ml) distances for the inserted element against the ml distances of the extein sequence.

Phages CactusJack, Glaske, Phalm, StressBall, Megiddo, KilKor, and Willsammy were isolated in Texas; they group together in both the extein and intein phylogenies (Figs. 6A and B). The inteins in these phages are identical (Fig. 6B), whereas the extein sequences show divergence (Fig. 6A). Similarly, phages TootsiePop and Misha28 were isolated from Massachusetts, and Awesomesause from Rhode Island. These phages group together in both the intein and extein phylogenies; however, the inteins are identical, whereas the exteins show minor sequence divergence.

4. Discussion

The family of putative DNA methylases on which we report here has a sporadic distribution in the clusters of Actinophages (Fig. 1). These methylases are surprisingly divergent for a gene invaded by an intein [9]. Given this divergence, the reconstruction of the evolutionary history of these methylases must be considered with caution. The homologous methylases from phages belonging to the same cluster (clusters A, AS, AR, E, F, I, J) do not group together (Fig. 5 and supplementary Figure 1A and B). However, statistical tests provide strong support only for the sequences from cluster F to not form a clan (Table 2). Nevertheless, the finding that the F-cluster sequences do not group together, and the fact that only a fraction of genomes in each cluster contain a homolog to this methylase, suggest that these methylases were frequently gained and lost by the phages. This observation is similar to the studies of RMSs in bacteria [22] and archaea [23], which found that RMSs often are part of the mobilome, are gained though horizontal gene transfer, and are not fixed in a lineage.

The intein we investigated contains a homing endonuclease domain of the LAGLIDADG type. The second endonuclease we discovered is also a LAGLIDADG type endonuclease but is located outside the intein. It is found in a group of related phages from the F cluster (Figs 4, 5, 6D). The sequence divergence for this endonuclease, with exception of the sequence from phage Kersh, correlates reasonably well with the divergence of the extein. The significant similarity between this domain and a homing endonuclease from a group I intron suggests that this domain might represent an independent selfish genetic element that targets a region upstream of the intein insertion site; however, we do not find strong evidence for this domain to have been transferred between phage lineages (Table 2, Fig. 5, Figs. 6D and E). An alternative explanation is that the endonuclease domain is part of an RMS that contains both the endonuclease and methylation activity in the same peptide [25]. If this were the case, then the presence of the second endonuclease might represent the original form of the enzyme with the second endonuclease domain being lost from most sequences. However, the latter explanation is unlikely as LAGLIDADG endonucleases are known for their long recognition sites, function in homing, and have not been described as part of RMSs. In the structure predicted for the Taj methylase (Fig. 4B) the second homing endonuclease forms its own domain, and the remainder of the structure is similar to the predicted structure of the Poptart methylase (Fig. 4A). This suggests that the presence of this homing endonuclease domain might not interfere with the function of the methylase.

The inteins in the methylases fall into two well supported types (blue and red labels in Fig. 6B). The observation that the methylases which group together in the extein phylogeny (e.g., Phages Lilac and Hannaconda, or Martik and Shilan), harbor two different intein types reveals that the inteins jumped between divergent host proteins. The transfer of inteins between divergent phages is also illustrated by intein containing phages Hannaconda and Lilac, whose methylases are placed in well supported groups that otherwise do not contain inteins (Fig 5.). However, even in instances where several of the intein containing methylases group together and are invaded by the same type of intein, a closer inspection suggests likely recent transfer of the intein. Ignoring branch lengths, one could assume that a single intein invasion occurred at the base of the seven intein containing

 doi:10.20944/preprints202212.0517.v1

phages that were isolated in Texas (CactusJack, Glaske, Phalm, StressBall, Megiddo, KilKor, and Willsammy; Fig. 5). However, the methylase sequences have significantly diverged – in the compact alignment by over .7 substitutions per site on average (Figure 6A); whereas the intein sequences are identical. This suggests that the inteins recently spread among the phages isolated in Texas, a long time after their methylases had diverged.

Whereas the intein and the second endonuclease encode a homing endonuclease domain that likely facilitates the invasion of alleles with an empty target site, the disjunct distribution of the ShiLan domain remains enigmatic. Nevertheless, the ShiLan domain is found in divergent methylases (Fig. 5) and, similar to each of the two intein types, the ShiLan domains, with one exception, have diverged much less than the associated methylase sequences. This suggests that the ShiLan domain too were transferred between divergent methylases. This notion is also supported by the AU-test (Table 2) which strongly rejects the hypothesis that the ShiLan domain containing methylases might from a coherent group in the methylase phylogeny.

5. Conclusions

Recombination between viruses has long been recognized as an important process. Even before the recognition of DNA as genetic material [39] Luria had inferred recombination between phages from multiplicity reactivation [40]. Despite its prominent role in the history of molecular biology, the amount of naturally occurring recombination we find in our study may be surprising to most. The intein and ShiLan domain distributions and phylogenies reveal frequent within gene recombination events between phages belonging to different clusters. In addition, the gene targeted for invasion has a sporadic distribution suggesting frequent gene loss and transfer events within and between phage clusters. Furthermore, the lack of divergence of the insertion element suggests local recent invasion of related phages by the intein.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: phylogenies calculated from the compact (A) and gappy (B) alignments. Figure S2: reliability score mapped on the methylase structure. PDB and python files (can be opened in chimera (v1.16)): PopTart_63_AF_predicted_structure.py, Taj_79_AF_predicted_structure.py, PopTart_63_AF_predicted_structure.pdb, Taj_79_AF_predicted_structure.pdb.

Author Contributions: Conceptualization, SPG and JPG; methodology, SPG and JPG; software, SPG and DRA.; validation, SPG and CAJ; formal analysis, CAJ, JPG; investigation SPG, CAJ and JPG.; resources, SPG, DRA, JPG; data curation, SPG and JPG; writing—original draft preparation JPG; writing—review and editing, SPG, JPG; visualization DRA; supervision, SPG and JPG; project administration, SPG and JPG; funding acquisition, JPG. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation within the BSF-NSF joint research program, NSF/MCB 1716046.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: All sequence data are available at PhagesDB https://phagesdb.org .

Acknowledgments: The Computational Biology Core of the Institute for Systems Genomics at the University of Connecticut provided computational resources. Resources and support were provided through the SEA-PHAGES program, in particular by Dan Russell from the Hatfull Lab, University of Pittsburgh. JPG thanks L. Thiberio Rangel for helpful discussions and suggestions. We thank the students in MCB 1201 Virus Hunting at the University of Connecticut for their shared fascination with inteins.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. SEA-PHAGES | Home Available online: https://seaphages.org/ (accessed on 4 December 2022).
2. Liu, X.Q. Protein-Splicing Intein: Genetic Mobility, Origin, and Evolution. *Annu Rev Genet* **2000**, *34*, 61–76, doi:10.1146/annurev.genet.34.1.61.
3. Pietrokovski, S. Intein Spread and Extinction in Evolution. *Trends Genet* **2001**, *17*, 465–472, doi:10.1016/S0168-9525(01)02365-4.
4. Perler, F.B.; Olsen, G.J.; Adam, E. Compilation and Analysis of Intein Sequences. *Nucleic Acids Res* **1997**, *25*, 1087–1093, doi:10.1093/NAR/25.6.1087.
5. Gogarten, J.P.; Senejani, A.G.; Zhaxybayeva, O.; Olendzenski, L.; Hilario, E. Inteins: Structure, Function, and Evolution. *Annu Rev Microbiol* **2002**, *56*, 263–287, doi:10.1146/annurev.micro.56.012302.160741.
6. Perler, F.B. Protein Splicing of Inteins and Hedgehog Autoproteolysis: Structure, Function, and Evolution. *Cell* **1998**, *92*, 1–4, doi:10.1016/S0092-8674(00)80892-2.
7. Gimble, F.S.; Thorner, J. Homing of a DNA Endonuclease Gene by Meiotic Gene Conversion in Saccharomyces Cerevisiae. *Nature* **1992**, *357*, 301–306, doi:10.1038/357301A0.
8. Naor, A.; Altman-Price, N.; Soucy, S.M.; Green, A.G.; Mitiagina, Y.; Turgeman-Grotta, I.; Davidovich, N.; Gogarten, J.P.; Gophna, U. Impact of a Homing Intein on Recombination Frequency and Organismal Fitness. *Proc Natl Acad Sci U S A* **2016**, *113*, E4654–E4661, doi:10.1073/pnas.1606416113.
9. Swithers, K.S.; Senejani, A.G.; Fournier, G.P.; Gogarten, J.P. Conservation of Intron and Intein Insertion Sites: Implications for Life Histories of Parasitic Genetic Elements. *BMC Evol Biol* **2009**, *9*, doi:10.1186/1471-2148-9-303.
10. Kelley, D.S.; Lennon, C.W.; Belfort, M.; Novikova, O. Mycobacteriophages as Incubators for Intein Dissemination and Evolution. *mBio* **2016**, *7*, doi:10.1128/MBIO.01537-16.
11. Barka, E.A.; Vatsa, P.; Sanchez, L.; Gaveau-Vaillant, N.; Jacquard, C.; Klenk, H.-P.; Clément, C.; Ouhdouch, Y.; van Wezel, G.P. Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev* **2015**, *80*, 1–43, doi:10.1128/MMBR.00019-15.
12. Guerrero-Bustamante, C.A.; Dedrick, R.M.; Garlena, R.A.; Russell, D.A.; Hatfull, G.F. Toward a Phage Cocktail for Tuberculosis: Susceptibility and Tuberculocidal Action of Mycobacteriophages against Diverse Mycobacterium Tuberculosis Strains. *mBio* **2021**, *12*, doi:10.1128/MBIO.00973-21.
13. Hatfull, G.F.; Dedrick, R.M.; Schooley, R.T. Phage Therapy for Antibiotic-Resistant Bacterial Infections. *Annu Rev Med* **2022**, *73*, 197–211, doi:10.1146/ANNUREV-MED-080219-122208.
14. Fruciano, E.; Bourne, S. Phage as an Antimicrobial Agent: D'Herelle's Heretical Theories and Their Role in the Decline of Phage Prophylaxis in the West. *The Canadian Journal of Infectious Diseases & Medical Microbiology* **2007**, *18*, 19, doi:10.1155/2007/976850.
15. Russell, D.A.; Hatfull, G.F. PhagesDB: The Actinobacteriophage Database. *Bioinformatics* **2017**, *33*, 784–786, doi:10.1093/BIOINFORMATICS/BTW711.
16. The Actinobacteriophage Database | Home Available online: https://phagesdb.org/ (accessed on 6 December 2022).
17. Pope, W.H.; Bowman, C.A.; Russell, D.A.; Jacobs-Sera, D.; Asai, D.J.; Cresawn, S.G.; Jacobs, W.R.; Hendrix, R.W.; Lawrence, J.G.; Hatfull, G.F. Whole Genome Comparison of a Large Collection of Mycobacteriophages Reveals a Continuum of Phage Genetic Diversity. *Elife* **2015**, *4*, doi:10.7554/ELIFE.06416.
18. Løbner-Olesen, A.; Skovgaard, O.; Marinus, M.G. Dam Methylation: Coordinating Cellular Processes. *Curr Opin Microbiol* **2005**, *8*, 154–160, doi:10.1016/J.MIB.2005.02.009.
19. Katayama, T. Initiation of DNA Replication at the Chromosomal Origin of E. Coli, OriC. *Adv Exp Med Biol* **2017**, *1042*, 79–98, doi:10.1007/978-981-10-6955-0_4.

20. Mruk, I.; Kobayashi, I. To Be or Not to Be: Regulation of Restriction-Modification Systems and Other Toxin-Antitoxin Systems. *Nucleic Acids Res* **2014**, *42*, 70–86, doi:10.1093/NAR/GKT711.
21. Kobayashi, I. Behavior of Restriction-Modification Systems as Selfish Mobile Elements and Their Impact on Genome Evolution. *Nucleic Acids Res* **2001**, *29*, 3742–3756.
22. Kong, Y.; Ma, J.H.; Warren, K.; Tsang, R.S.W.; Low, D.E.; Jamieson, F.B.; Alexander, D.C.; Hao, W. Homologous Recombination Drives Both Sequence Diversity and Gene Content Variation in Neisseria Meningitidis. *Genome Biol Evol* **2013**, *5*, 1611–1627, doi:10.1093/gbe/evt116.
23. Fullmer, M.S.; Ouellette, M.; Louyakis, A.S.; Papke, R.T.; Gogarten, J.P. The Patchy Distribution of Restriction–Modification System Genes and the Conservation of Orphan Methyltransferases in Halobacteria. *Genes (Basel)* **2019**, *10*, 233, doi:10.3390/genes10030233.
24. Wilson, G.G.; Murray, N.E. Restriction and Modification Systems. *Annu Rev Genet* **1991**, *25*, 585–627, doi:10.1146/ANNUREV.GE.25.120191.003101.
25. Smith, R.M.; Pernstich, C.; Halford, S.E. TstI, a Type II Restriction-Modification Protein with DNA Recognition, Cleavage and Methylation Functions in a Single Polypeptide. *Nucleic Acids Res* **2014**, *42*, 5809–5822, doi:10.1093/NAR/GKU187.
26. Cresawn, S.G.; Bogel, M.; Day, N.; Jacobs-Sera, D.; Hendrix, R.W.; Hatfull, G.F. Phamerator: A Bioinformatic Tool for Comparative Bacteriophage Genomics. *BMC Bioinformatics* **2011**, *12*, 395, doi:10.1186/1471-2105-12-395.
27. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res* **1997**, *25*, 3389–3402, doi:10.1093/nar/25.17.3389.
28. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **2013**, *30*, 772–780, doi:10.1093/MOLBEV/MST010.
29. Gouy, M.; Tannier, E.; Comte, N.; Parsons, D.P. Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation. *Methods Mol Biol* **2021**, *2231*, 241–260, doi:10.1007/978-1-0716-1036-7_15.
30. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R.; Teeling, E. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **2020**, *37*, 1530, doi:10.1093/MOLBEV/MSAA015.
31. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589, doi:10.1038/S41586-021-03819-2.
32. Meng, E.C.; Pettersen, E.F.; Couch, G.S.; Huang, C.C.; Ferrin, T.E. Tools for Integrated Sequence-Structure Analysis with UCSF Chimera. *BMC Bioinformatics* **2006**, *7*, 339, doi:10.1186/1471-2105-7-339.
33. Zimmermann, L.; Stephens, A.; Nam, S.-Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A.N.; Alva, V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at Its Core. *J Mol Biol* **2017**, doi:10.1016/J.JMB.2017.12.007.
34. UVA FASTA Server Available online: https://fastademo.bioch.virginia.edu/fasta_www2/fasta_list2.shtml (accessed on 7 December 2022).
35. Kishino, H.; Hasegawa, M. Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea. *J Mol Evol* **1989**, *29*, 170–179, doi:10.1007/BF02100115.
36. Shimodaira, H.; Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* **1999**, *16*, 1114–1114, doi:10.1093/OXFORDJOURNALS.MOLBEV.A026201.

 doi:10.20944/preprints202212.0517.v1

37. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst Biol* **2002**, *51*, 492–508, doi:10.1080/10635150290069913.
38. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **2015**, *32*, 268–274, doi:10.1093/molbev/msu300.
39. HERSHEY, A.D.; CHASE, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J Gen Physiol* **1952**, *36*, 39–56, doi:10.1085/JGP.36.1.39.
40. Luria, S.E. Reactivation of Irradiated Bacteriophage by Transfer of Self-Reproducing Units. *Proc Natl Acad Sci U S A* **1947**, *33*, 253–264, doi:10.1073/PNAS.33.9.253.