Submitted to genes

MDPI

Article

The Patchy Distribution of Restriction-Modification System Genes and the Conservation of Orphan Methyltransferases in Halobacteria

Matthew S. Fullmer[1,2,$], Matthew Ouellette[1,$], Artemis S. Louyakis[1,$], R. Thane Papke[1*], J. Peter Gogarten[1*]

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[2] Present Address: Bioinformatics Institute, School of Biological Sciences, The University of Auckland, Auckland, New Zealand
[$] These authors contributed equally
[*] Correspondence: thane@uconn.edu; gogarten@uconn.edu

Received: date; Accepted: date; Published: date

Abstract: Restriction-modification (RM) systems in Bacteria are implicated in multiple biological roles ranging from defense against parasitic genetic elements, to selfish addiction cassettes, and barriers to gene transfer and lineage homogenization. In Bacteria, DNA-methylation without cognate restriction also plays important roles in DNA replication, mismatch repair, protein expression, and in in biasing DNA uptake. Little is known about archaeal RM systems and DNA methylation. To elucidate further understanding for the role of RM systems and DNA methylation in Archaea, we undertook a survey of the presence of RM system genes and related genes, including orphan DNA methylases, in the halophilic archaeal class Halobacteria. Our results reveal that some orphan DNA methyltransferase genes were highly conserved among lineages indicating an important functional constraint, whereas RM systems demonstrated patchy patterns of presence and absence. This irregular distribution is due to frequent horizontal gene transfer and gene loss, a finding suggesting that the evolution and life cycle of RM systems may be best described as that of a selfish genetic element. A putative target motif (CTAG) of one of the orphan methylases was underrepresented in all of the analyzed genomes, whereas another motif (GATC) was overrepresented in most of the haloarchaeal genomes, particularly in those that encoded the cognate orphan methylase.

Keywords: HGT, Restriction, Methylation, Gene Transfer, Selfish Genes, Archaea, Haloarchaea, DNA methylase, epigenetics

Funding: This work was supported through grants from the Binational Science Foundation (BSF 2013061); the National Science Foundation (NSF/MCB 1716046) within the BSF-NSF joint research program; and NASA exobiology (NNX15AM09G, and 80NSSC18K1533).

1. Introduction

DNA methyltransferases (MTases) are enzymes which catalyze the addition of a methyl group to a nucleotide base in a DNA molecule. These enzymes will methylate either adenine, producing *N*6-methyladenine (6mA), or cytosine, producing either *N*4-methylcytosine (4mC) or *C*5-methylcytosine (5mC), depending on the type of MTase enzyme [1]. DNA methyltransferases typically consist of three types of protein domains: an S-adenosyl-L-methionine (AdoMet) binding domain which obtains the methyl group from the co-factor AdoMet, a target recognition domain (TRD) which binds the enzyme to the DNA strand at a short nucleotide sequence known as the recognition sequence, and a catalytic domain which transfers the methyl group from AdoMet to a

nucleotide at the recognition sequence [2]. The order in which these domains occur in an MTase varies and can be used to classify the enzymes into the subtypes of α, β, γ, δ, ε, and ζ MTases [3–5].

In bacteria and archaea, MTases are often components of restriction-modification (RM) systems, in which an MTase works alongside a cognate restriction endonuclease (REase) that targets the same recognition site. The REase will cleave the recognition site when it is unmethylated, but the DNA will escape cutting when the site has been methylated by the MTase; this provides a self-recognition system to the host where it differentiates between its own methylated DNA and that of unmethylated, potentially harmful foreign DNA that is then digested by the host's REase [6–8]. RM systems have also been described as addiction cassettes akin to toxin-antitoxin systems, in which post-segregational killing occurs when the RM system is lost since the MTase activity degrades more quickly than REase activity, resulting in digestion of the host genome at unmodified recognition sites [9,10]. RM systems have been hypothesized to act as barriers to genetic exchange and drive population diversification [11,12]. In *Escherichia coli*, for example, conjugational uptake of plasmids is reduced by the RM system EcoKI when the plasmids contain EcoKI recognition sequences [13]. However, transferred DNA that is digested by a cell's restriction endonuclease can still effectively recombine with the recipient's chromosomal DNA [7,14,15]; the effect of DNA digestion serves to limit homologous recombinant DNA fragment size [16]. Restriction thus advantages its host by decreasing transfer of large mobile genetic elements and infection with phage originating in organisms without the cognate MTase [8], while also reducing linkage between beneficial and slightly deleterious mutations [17].

There are four major types of RM systems which have been classified in bacteria and archaea [18,19]. Type I RM systems consist of three types of subunits: REase (R) subunits, MTase (M) subunits, and site specificity (S) subunits which contain two tandem TRDs. These subunits form pentamer complexes of two R subunits, two M subunits, and one S subunit, and these complexes will either fully methylate recognition sites which are modified on only one DNA strand (hemimethylated) or cleave the DNA several bases upstream or downstream of recognition sites which are unmethylated on both strands [20,21]. The MTases and REases of Type II RM systems have their own TRDs and operate independently of each other, but each one targets the same recognition site [22]. There are many different subclasses of Type II RM system enzymes, such as Type IIG enzymes which contain both REase and MTase domains and are, therefore, capable of both methylation and endonuclease activity [23]. Type III RM systems consist of REase (Res) and MTase (Mod) subunits which work together as complexes, with the Mod subunit containing the TRD which recognizes asymmetric target sequences [24]. Type IV RM systems are made up of only REases, but unlike in other RM systems, these REases will target and cleave methylated recognition sites [20,25].

MTases can also exist in bacterial and archaeal hosts as orphan MTases, in which they occur independently of cognate restriction enzymes and typically have important physiological functions [26]. In *E. coli*, the orphan MTase Dam, an adenine MTase which targets the recognition sequence GATC, is involved in regulating the timing of DNA replication by methylating the GATC sites present at the origin of replication (*oriC*) [27]. The protein SeqA binds to hemimethylated GATC sites at *oriC*, which prevents re-initiation of DNA replication at *oriC* after a new strand has been synthesized [28,29]. Dam methylation is also important in DNA repair in *E. coli*, where the methylation state of GATC sites is used by the methyl-directed mismatch repair (MMR) system to identify the original DNA strand in order to make repairs to the newly-synthesized strand [30–32]. In *Cauldobacter crescentus*, the methylation of target sites in genes such as *ctrA* by orphan adenine MTase CcrM helps regulate the cell cycle of the organism [33–35]. The importance of orphan MTases in cellular processes is likely the reason why they are more widespread and conserved in bacteria compared to MTases associated with RM systems [36,37].

MTases and RM systems have been well-studied in the bacteria, but less research has been performed in archaea, with most studies focused on characterizing RM systems of thermophilic species [38–42]. Recent research into the halophilic archaeal species *Haloferax volcanii* has demonstrated a role for DNA methylation in DNA metabolism, and probably uptake: cells could not grow on wild type *E. coli* DNA as a phosphorous source, whereas unmethylated *E. coli* was metabolized completely [43,44]. In an effort to better understand this phenomenon, we characterized the genomic methylation patterns (methylome) and MTases in the halophilic archaeal species *Haloferax volcanii* [45,46]. However, the distribution of RM systems and MTases among the archaea has not been extensively studied, and thus their life histories and impact on host evolution are unclear.

To that end we surveyed the breadth of available genomes from public databases representing the class Halobacteria, also known as the Haloarchaea, for RM system and MTase candidate genes. We further sequenced additional genomes from the genus *Halorubrum* which provided an opportunity to examine patterns among very closely related strains. Upon examining their patterns of occurrence, we discovered orphan methyltransferases widely distributed throughout the Haloarchaea. In contrast, RM system candidate genes had a sparse and spotty distribution indicating frequent gene transfer and loss. Even individuals from the same species isolated from the same environment and at the same time, differed in the RM system complement.

2. Materials and Methods

Search Approach. The starting data consists of 217 Halobacteria genomes from NCBI and 14 in-house sequenced genomes (Supplementary Table S1**)**. We note that some of these genomes were assembled from shotgun metagenome sequences and not from individual cultured strains. Genome completion was determined through identification of 371 Halobacteriaceae marker genes using CheckM v1.0.7 [47]. Queries for all restriction-methylation-specificity genes were obtained from the Restriction Enzyme dataBASE (REBASE) website [48,49]. As methylation genes are classified by function rather than by homology [48] the protein sequences of each category were clustered into homologous groups (HGs) via the *uclust* function of the USEARCH v9.0.2132 package [50] at a 40 percent identity. The resulting ~36,000 HGs were aligned with MUSCLE v3.8.31 [51]. HMMs were then generated from the alignments using the *hmmbuild* function of HMMER3 v3.1b2 (hmmer.org). The ORFs of the 217 genomes were searched against the profiles via the *hmmsearch* function of HMMER3. Top hits were extracted and cross hits filtered with in-house Perl scripts. Steps were taken to collapse and filter HGs. First, the hits were searched against the arCOG database [52] using BLAST [53] to assign arCOG identifiers to the members of each group. Second the R package *igraph* v1.2.2 [54] was used to create a list of connected components from the arCOG identifications. All members of a connected component were collapsed into a single collapsed HG (cHG).

Because REBASE is a database of all methylation-restriction-related activities there are many members of the database outside our interest. At this point we made a manual curation of our cHGs attempting to identify known functions that did not apply to our area of interest. Examples include protein methylation enzymes, exonucleases, cell-division proteins, etc. The final tally of this clustering and filtering yielded 1696 hits across 48 total candidate cHGs. arCOG annotations indicate DNA methylase activity, restriction enzyme activity, or specificity module activity as part of an RM system for 26 cHGs. The remaining 22 cHGs had predominant arCOG annotations matching other functions that may reasonably be excluded from conservative RM system-specific analyses. For a graphical representation of the search strategy see supplementary materials **Figure S1.**

Reference Phylogeny. A reference tree was created using the full complement of ribosomal proteins. The ribosomal protein set for *Halorubrum lacusprofundi* ATCC 49239 was obtained from the BioCyc website [55]. Each protein orf was used as the query in a BLAST [53] search against each genome. Hits for each gene were aligned with MUSCLE v3.8.31 [51] and then concatenated with in-

house scripting. The concatenated alignment was subjected to maximum likelihood phylogenetic inference in the IQ-TREE v1.6.1 suite with ultrafast bootstrapping and automated model selection [56,57]. The final model selection was LG+F+R9.

F81 Presence-Absence Phylogeny. It is desirable to use maximum-likelihood methodology rather than simple distance measures. To realize this, the matrix was converted to an A/T alignment by replacing each present with an "A" and absent with a "T." This allowed use of an F81 model with empirical base frequencies. This confines the base parameters to only A and T while allowing all of the other advantages of an ML approach. IQ-TREE was employed to infer the tree with 100 bootstraps [57].

Horizontal Gene Transfer Detection. Gene trees for each of the cHGs were inferred using RAxML v8.2.11 [58] under PROTCATLG models with 100 bootstraps. The gene trees were then improved by resolving their poorly supported in nodes to match the species tree using TreeFix-DTL [59]. Optimized gene tree rootings were inferred with the OptRoot function of Ranger-DTL. Reconciliation costs for each gene tree were computed against the reference tree using Ranger-DTL 2.0 (http://compbio.engr.uconn.edu/software/RANGER-DTL/) [60] with default DTL costs. One-hundred reconciliations, each using a different random seed, were calculated for each cHG. After aggregating these with the AggregateRanger function of Ranger-DTL the results were summarized and each prediction and any transfer inferred in 51% or greater of cases was counted as a transfer event.

Data Analysis and Presentation: The presence-absence matrix of cHGs was plotted as a heatmap onto the reference phylogeny using the *gheatmap* function of the R Bioconductor package *ggtree* v1.14.4 [61,62]. The rarefaction curve was generated with the *specaccum* function of the *vegan* v2.5-3 package in R [63] and number of genomes per homologous group was plotted with ggplot2 v3.1.0 [64]. Spearman correlations and significances between the presence-absence of cHGs was calculated with the *rcorr* function of the *hmisc* v4.1-1 package in R (http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc). A significance cutoff of $p < 0.05$ was used with a Bonferroni correction. All comparisons failing this criterion were set to correlation = 0. These data were plotted into a correlogram via the *corrplot* function of the R package *corrplot* v0.84. To compare the Phylogeny calculated from Presence-Absence data to the ribosomal protein reference, the bootstrap support set of the presence-absence phylogeny was mapped onto the ribosomal protein reference tree using the *plotBS* function in *phangorn* v2.4.0 [65]. Support values equal to or greater than 10% are displayed. To compare phylogenies using Internode Certainty, scores were calculated using the IC/TC score calculation algorithm implemented in RAxML v8.2.11 [58,66].

Synteny. Genomes were searched for location of cHGs. Proximity was used to determine synteny of groups of cHGs frequently identified on the same genomes.

Presence-Absence PCoA. Jaccard distances between presence-absence of taxa were calculated using the *distance* function of the R package *philentropy* v0.2.0 [67]. The PCoA was generated using the *wcmdscale* function in *vegan* v2.5-3 [63]. The two best sampled genera, *Halorubrum* (orange) and *Haloferax* (red), are colored distinctively.

Recognition Site Assignment. To determine the most likely recognition sites, each member of each cHG was searched against the REBASE Gold Standard set using BLASTp. The REBASE gold standard set was chosen over the individual gene sets on account of it having a much higher density of recognition site annotation. This simplifies the need to search for secondary hits to find predicted target sites. After applying an e-value cut-off of 1E-20, the top hit was assigned to each ORF.

CTAG and GATC motifs were counted with an inhouse perl script available at the Gogarten-lab's GitHub [68].

Gene Ontology. Sets of GO terms were identified for each cHG using Blast2GO [69]. Annotations were checked against the UniProt database [70] using arCOG identifiers.

3. Results

RM-system gene distribution

Analysis of 217 haloarchaeal genomes and metagenome assembled genomes yielded 48 total candidate collapsed homologous groups (cHGs) of RM-system components. Out of these 48 cHGs, 26 had arCOG annotation suggesting DNA methylase activity, restriction enzyme activity, or specificity module activity as part of an RM system. We detected 22 weaker candidates with predominant arCOG annotations matching other functions (**Table 1)**. Our analysis shows that nearly all of the cHGs are found more than once. (**Figure 1A)**. Indeed, 16 families are found in 20 or more genomes each (>9 %), and this frequency steadily increases culminating in five families being conserved in greater than 80 genomes each (>37 %) with one cHG being in ~80 % of all Haloarchaea surveyed. Though these genes appear frequently in taxa across the haloarchaeal class, the majority of each candidate RM system cHG is present in fewer than half the genomes, - the second most abundantly recovered cHG is found in only ~47 % of all taxa surveyed. We note that the cHGs with wide distribution are annotated as MTases without an identifiable co-evolving restriction endonuclease: Group U DNA_methylase-022; W dam_methylase-031; Y dcm_methylase-044; and AT Uncharacterized-032 (members of this cHG are also annotated as methylation subunit and N6-Adenine MTase). Rarefaction analysis indicates about 50 % of the genomes assayed contain seven dominant cHGs, and that all taxa on average are represented by half of the cHGs (**Figure 1B)**. Together, the separate analyses indicate extensive gene gain and loss of RM-system genes. In contrast, orphan MTases in cHG U and W, and to a lesser extent Y (Figure 2) have a wider distribution in some genera (see below for further discussion).

The phylogeny of the class Halobacteria inferred from concatenated ribosomal proteins (**Figure 2)** was largely comparable to prior work [71], and with a taxonomy based on concatenations of conserved proteins [72,73]. For instance, in our phylogeny the *Halorubracaea* group with the *Haloferacaceae* recapitulating the order *Haloferacales*, and the families, *Halobacteriaceae*, *Haloarculaceae* and *Halococcaceae* group within the order *Halobacteriales*. Our genome survey in search of RM-system genes encompassed a broad taxonomic sampling, and it explores in depth the genus *Halorubrum* because it is a highly speciated genus, and because the existence of many genomes from the same species allows within species distribution assessment.

Comparison of the phylogeny in **Figure 2** to the heatmap giving the presence/absence of RM system cHG candidates demonstrates that the cHG distribution is highly variable (**Figure 2)**. The one glaring exception is cHG U, a DNA methylase found in 174 of the 217 genomes analyzed. Since it is not coupled with a restriction enzyme of equal abundance, it is assumed to be an orphan MTase. The MTase from *Hfx. volcanii* (gene HVO_0794), which recognizes the CTAG motif [45] is a member of this cHG. Though U is widely distributed, within the genus *Halorubrum* it is only found in ~37.5 % (21/56) of the genomes. While U's phylogenetic profile is compatible with vertical inheritance over much of the phylogeny, the presence absence data also indicate a few gene transfer and loss events within *Halorubrum*. cHG U is present in *Hrr. tebenquichense* DSM14210, *Hrr. hochstenium* ATCC700873, *Hrr.* sp. AJ767, and in strains from the related species *Hrr. distributum*, *Hrr. arcis*, *Hrr. litoreum* and *Hrr. terrestre* suggesting an acquisition in the ancestor of this group.

Instead of U, another orphan MTase is abundantly present in *Halorubrum* spp., cHG W. It was found in ~95 % of all *Halorubrum* strains, with three exceptions - an assembled genome from

metagenome sequence data, and two from incomplete draft genomes of the species *Halorubrum ezzemoulense*. Interestingly, when U is present in a *Halorubrum* sp. genome, so too is W **(Figure 2)**. In a complementary fashion, analysis of W outside of the *Halorubrum* shows that it is found patchily distributed throughout the rest of the class Halobacteria (~20 % -32/158), and always as a second orphan MTase with cHG U. When the members of cHG W were used to search the uniprot database, the significant matches included the *E. coli* Dam MTase, a very well-characterized GATC MTase, which provides strong evidence that this cHG is a GATC orphan MTase family. The presence and absence of cHG U and W in completely sequenced genomes is given in Table S3, together with the frequency of the CTAG and GATC motifs in the main chromosome.

The rest of the RM cHGs are much more patchily distributed (**Figure 2**). For instance, the cHGs that make up columns A-G represent different gene families within the Type I RM system classification; two MTases (A,B), three REases (C,D,E), and two site specificity units (SSUs) (F,G). Throughout the Haloarchaea, cHGs from columns A, E and F, representing an MTase, an REase, and an SSU respectively, are found co-occurring 35 times. In a subset of genomes studied for synteny A, E and F are encoded next to one another in *Natrinema gari, Halorhabdus utahensis*, *Halorubrum* SD690R, *Halorubrum ezzemoulense* G37, and *Haloorientalis* IM1011 (**Figure 3**). These genes probably represent a single transcriptional unit of genes working together for restriction and modification purposes. Since the Type I RM system is a five-component system, the likely stoichiometry is 2:2:1. These three cHGs co-occur four times within the species *Halorubrum ezzemoulense*, and two of these cHGs (A and E) co-occur an additional three more times, suggesting either a loss of the SSU, or an incomplete genome sequence for those strains. If it is due to incomplete sequencing, then 7/16 (43 %) of the *Hrr. ezzemoulense* genomes have this set of co-occurring genes, while half do not have an identified Type I system. This is particularly stunning since strains FB21, Ec15, G37 and Ga2p were all cultivated at the same time from the same sample, a hypersaline lake in Iran. Furthermore, one strain, Ga36, has a different identified Type I RM system composed of substituted cHGs A and E with B and D, respectively, while maintaining the same SSU. This suggests the same DNA motif may be recognized by the different cHGs and that these cHGs are therefore functionally interchangeable. Members of cHGs B, F, and D were found as likely co-transcribed units in *Halococcus salifodinae, Natronolimnobius aegyptiacus, Halorubrum kocurii, Haloarcula amylolytica* (**Figure 3**). In *Halorubrum* DL, and *Halovivax ruber* XH70, genomes that contained members from cHGs A, B, D, E, and F these genes were not found in a single unit, suggesting that they do not form a single RM system. Together, these analyses suggest this Type I RM system has a wide but sporadic distribution, that this RM system is not required for individual survival, and that functional substitutions occur for cHGs.

Type II RM systems contain an MTase and an REase that target the same motif but do not require an associated SSU because each enzyme has its own TRD. The Type II RM system cHGs are in columns H-L for the MTases, and M-P for the REases. Memberships to the Type II MTase cHGs are far more numerous in the Haloarchaea than their REase counterpart, as might be expected when witnessing decaying RM systems through the loss of the REase. The opposite result, more REases is a more difficult scenario because an unmethylated host genome would be subject to restriction by the remaining cognate REase (e.g., addiction cassettes). There are 14 "orphan" Type II REases in **Figure 2**, but their cognate MTase's absence could be explained by incomplete genome sequence data.

Type III RM systems have been identified in cHGs Q (MTase) and R and S (REases). Type III MTases and REases (cHGs Q and R) co-occur almost exclusively in the species *Halorubrum ezzemoulense,* our most highly represented taxon. Furthermore, these Type III RM systems are highly restricted in their distribution to that species, with cHGs co-occuring only twice more throughout the Haloarchaea, and with a different REase cHG (S); once in *Halorubrum arcis*, and another in *Halobacterium* D1. Orphan MTases occurred twice in cHG Q. Of particular interest is that closely related strains also cultivated from Lake Bidgol in Iran but which are in a different but closely related

Halorubrum species (e.g., Ea8, IB24, Hd13, Ea1, Eb13) do not have a Type III RM system, implying though exposed to the same halophilic viruses, they do not rely on this system for avoiding virus infection.

cHGs Z-AV are not sufficiently characterized to pinpoint of their role in DNA RM systems or as MTase. These cHGs likely include homing endonuclease or enzymes modifying nucleotide in RNA molecules; however, their function as orphan MTases or restriction endonucleases can at present not be excluded.

Horizontal Gene Transfer explains patchy distribution

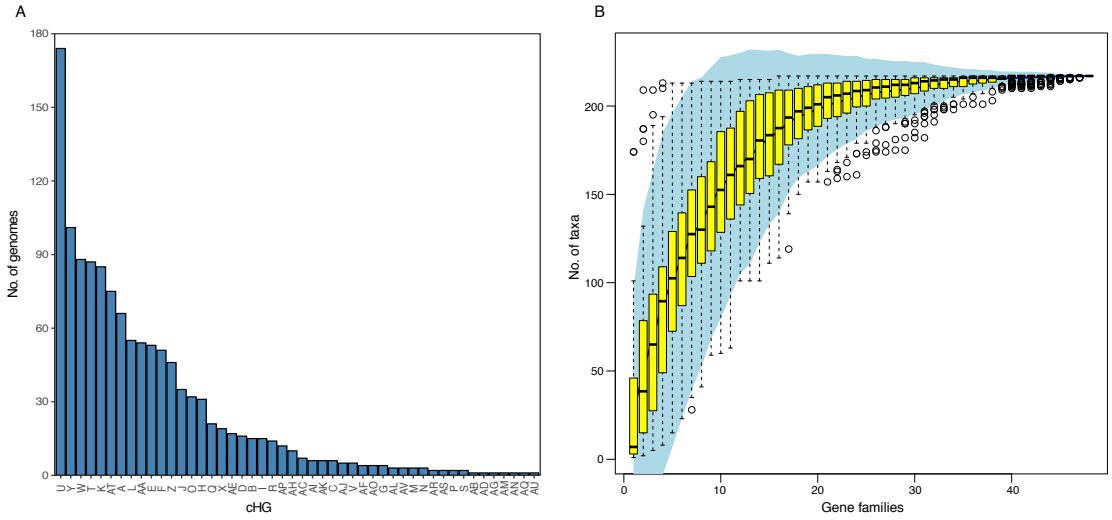
The patchy appearance of RM system candidates was further investigated by plotting the Jaccard distance of the presence-absence data against the alignment distance of the reference tree (supplementary **Figure S2)**. If the presence-absence data followed vertical descent one would expect the best-fit line to move from the origin with a strong positive slope. Instead, the best fit line is close to horizontal with an r-squared value of 0.0047, indicating negligible relationship between the overall genome phylogeny and RM system complement per genome. As another way of visualizing this data, the presence-absence patterns were plotted as a principle coordinates analysis **(**supplementary **Figure S3)**. The high degree of overlap between the ranges of the three groups illustrates that there are few RM system genes unique to a given group and a large amount of overlap in repertoires.

To further evaluate the the lack of long term vertical descent for RM system genes, a phylogeny was inferred from the presence-absence pattern of cHGs. The resultant tree (**Figure S4)** is largely in disagreement with the reference phylogeny. The bootstrap support set from the presence-absence phylogeny was mapped onto the ribosomal topology (**Figure S5)**. The resulting support values demonstrate an extremely small degree of agreement between the two methods. The few areas where there is even 10% support are near the tips of the ribosomal phylogeny and correspond to parts of established groups, such as *Haloferax*, *Natronobacterium*, and *Halorubrum*. Internode Certainty (IC) scores are another way to compare phylogenies. An average IC score of 1 represents complete agreement between the two phylogneies, and score of -1 complete disagreement. The average IC scores for the reference tree using the support set from the F81 tree was -0.509, illustrating that the presence absence data do not support the topology of the reference phylogeny.

The patchy distribution of the RM system candidate genes and their lack of conformity to the reference phylogeny suggests frequent horizontal gene transfer combined with gene loss events as the most probable explanation for the observed data. To quantify the amount of transfer the TreeFix-Ranger pipeline was employed. TreeFix-DTL resolves poorly supported areas of gene trees to better match the concatenated ribosomal protein gene tree used as reference. Ranger-DTL resolves optimal gene tree rooting against the species tree and then computes a reconciliation estimating the number of duplications, transfers, and losses that best explains the data (**Table 2)**. For almost every cHG with four or more taxa our analysis infers several HGT events. Only cHG R, a putative Type III restriction enzyme found only in a group of closely related *Halorubrum ezzemoulense strains*, has not been inferred to undergo at least one transfer event.

RM systems usually function as cooperative units [48,74,75]. It stands to reason that some of the RM system candidates may be transferred as units, maintaining their cognate functionality. This possibility was examined by a correlation analysis. A spearman correlation was made between all pairs of cHGs. Those with a significant result at a Bonferroni-corrected $p < 0.05$ were plotted in a correlogram (**Figure 4)**. As illustrated in **Figure 3**, cHGs with significant similar phylogenetic profiles often are near to one another in the genomes.

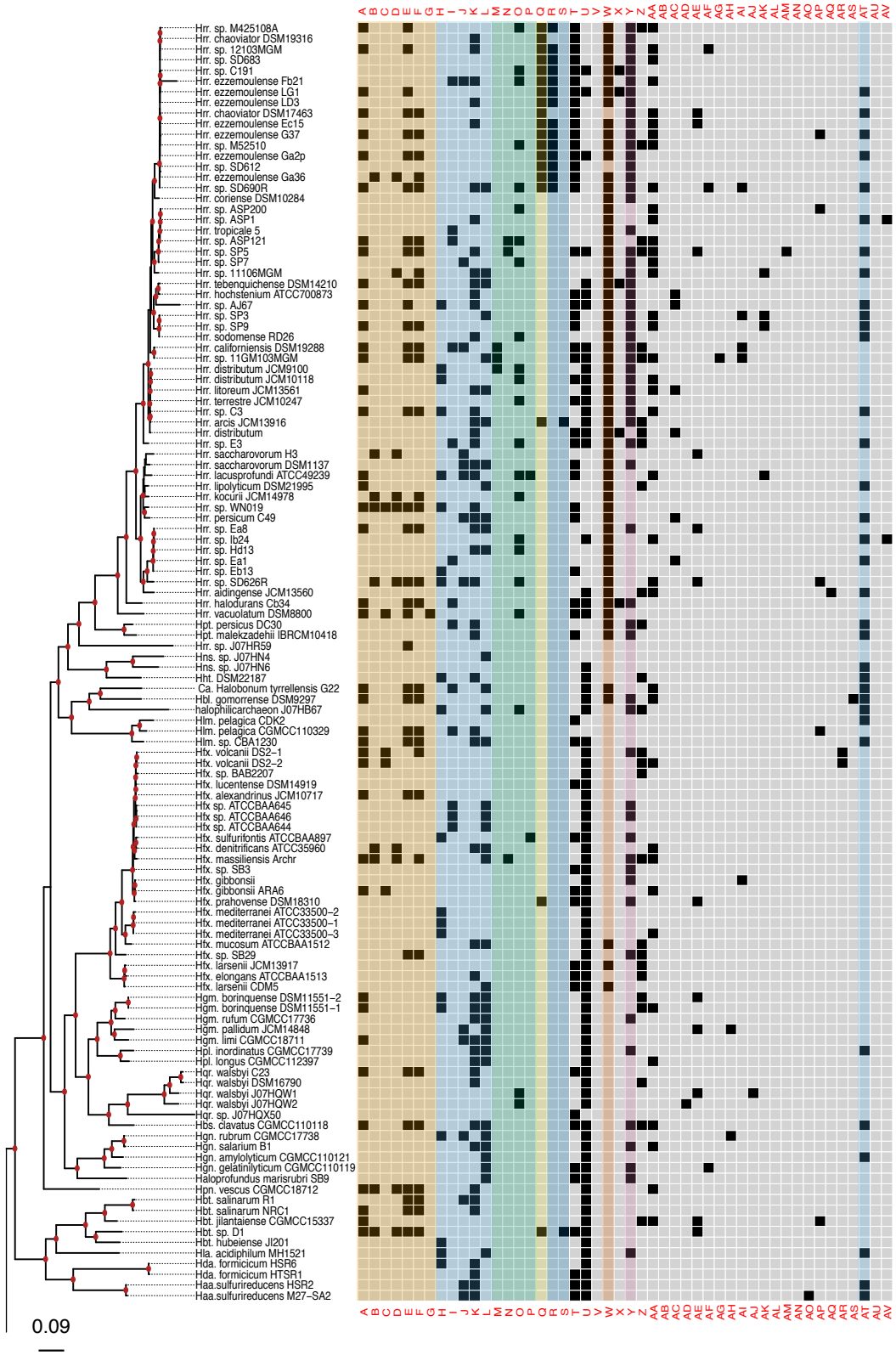
348 *3.2. Figures and Tables*



349

350 **Figure 1**: Distribution of collapsed Homologous Group (cHG) among haloarchaeal genomes. (A) the
351 number of genomes present in each collapsed Homologous Group (cHG). No cHG contains a
352 representative from every genome used in this study. With the exception of one cHG, all contain
353 members from fewer than half of the genomes. The cHGs are ordered by number of genomes they
354 contain. (B) rarefaction plot of the number of genomes represented as cHGs accumulate. 95%
355 confidence interval is shown in shaded blue area and yellow box whisker plots give the number of
356 taxa from random subsamples (permutations = 100) over 48 gene families.

357



358

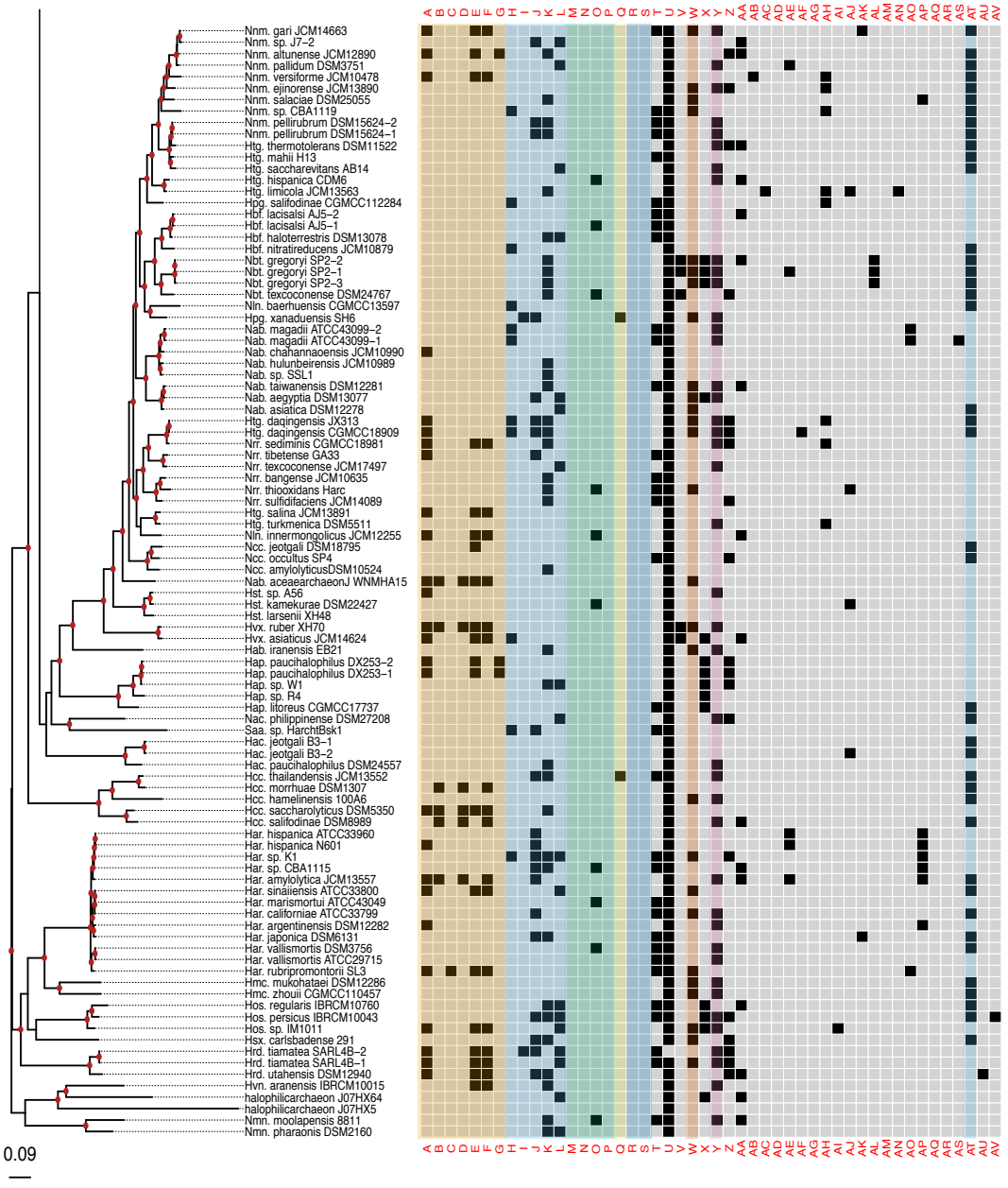
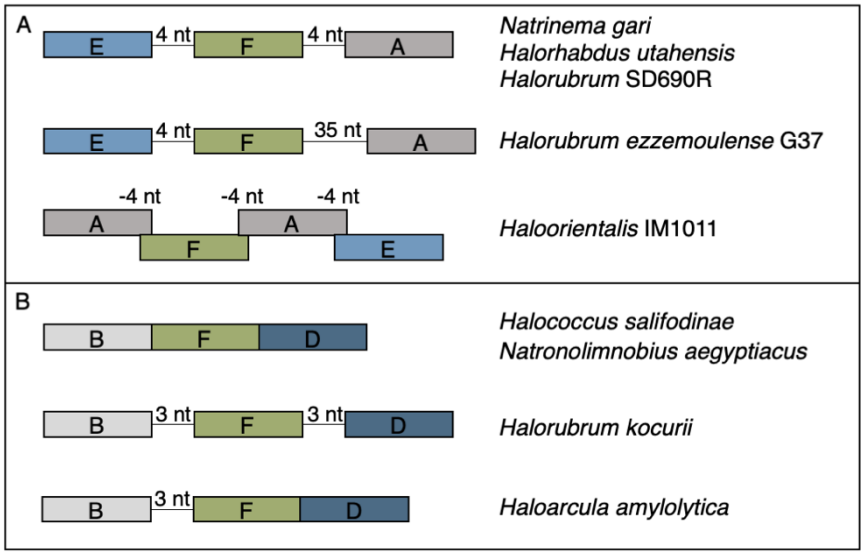
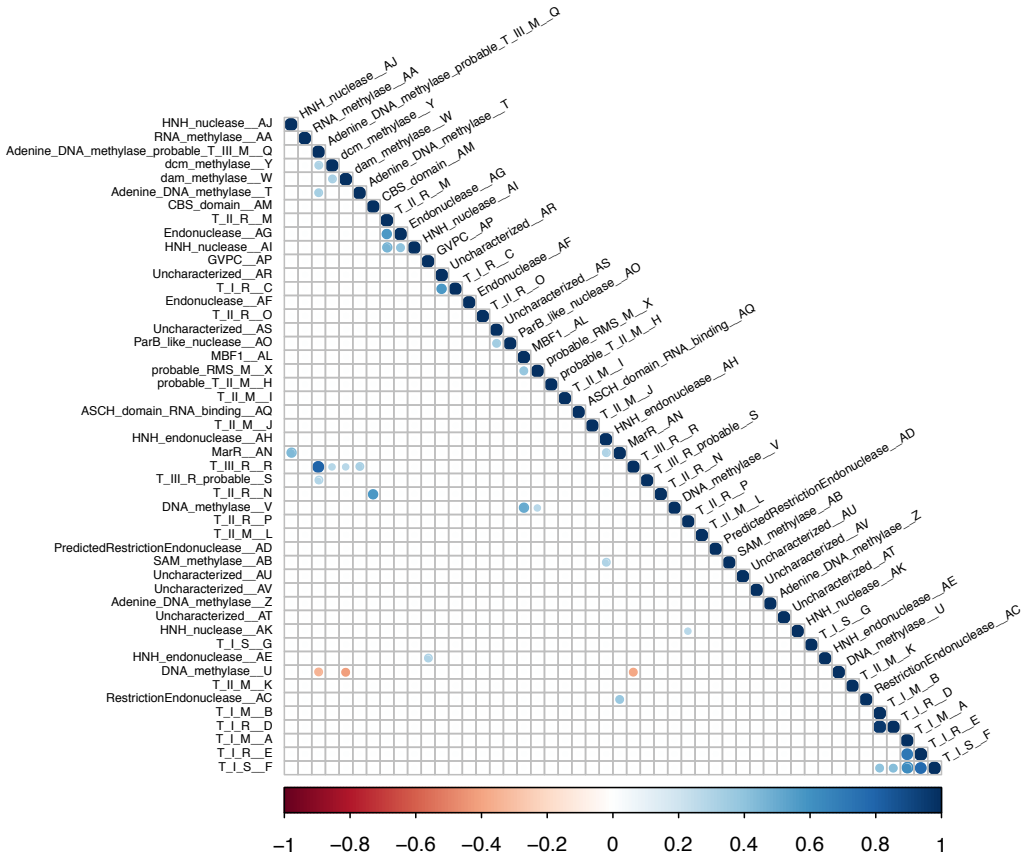


Figure 2: Presence-absence matrix of the 48 candidate RMS cHGs plotted against the reference phylogeny. For most cHGs the pattern of presence-absence does not match the reference phylogeny (compare supplementary Figures S2-S5) RMS-candidate cHGs are loosely ordered by system type and with the ambiguously assigned RM candidates at the end. Above is a key relating the column names to the majority functional annotation.



365
366 **Figure 3.** Gene maps for syntenic clusters of gene families (A) EFA and (B) BFD found in a subset of
367 organisms identified to the right of each map. Genes are colored by gene families with Type I
368 methylases (AB) in greys, Type I restriction endonucleases (DE) in blues, and Type I site specificity
369 unit (F) in green.



370

371 **Figure 4.** Heatmap of co-occurrence between the 48 RMS-candidate cHGs. Positive correlation
372 indicates the cHGs co-occur while negative indicates that the presence of one means the other will not
373 be present. Significance level is $p < 0.05$ with a Bonferroni correction applied for multiple tests. Blue
374 indicates significant positive correlation; red indicates a significant negative correlation.

375

Table 1. Collapsed homologous group descriptions. [$]

Alpha code	Numerical code	Annotated arCOG Function[$$]	arCOG number
A	cHG_021	T I M	arCOG02632
B	cHG_024	T I M	arCOG05282
C	cHG_018	T_I_R	arCOG00880
D	cHG_034	T I R	arCOG00879
E	cHG_045	T I R	arCOG00878
F	cHG_006	T I S	arCOG02626
G	cHG_025	T I S	arCOG02628
H	cHG_036	probable T II M	arCOG00890
I	cHG_001	T II M	arCOG02635
J	cHG_003	T II M	arCOG02634
K	cHG_011	T II M	arCOG04814
L	cHG_033	T II M	arCOG03521
M	cHG_007	T_II_R	arCOG11279
N	cHG_013	T II R	arCOG11717
O	cHG_023	T II R	arCOG03779
P	cHG_029	T_II_R	arCOG08993
Q	cHG_042	Adenine DNA methylase probable T III M	arCOG00108
R	cHG_008	T III R	arCOG06887
S	cHG_009	T III R probable	arCOG07494
T	cHG_014	Adenine DNA methylase	arCOG00889
U	cHG_022	DNA methylase	arCOG00115
V	cHG_027	DNA methylase	arCOG00129
W	cHG_031	dam methylase	arCOG03416
X	cHG_035	probable RMS M	arCOG08990
Y	cHG_044	dcm methylase	arCOG04157
Z	cHG_048	Adenine DNA methylase	arCOG02636
AA	cHG_010	RNA methylase	arCOG00910
AB	cHG_040	SAM-methylase	arCOG01792
AC	cHG_012	RestrictionEndonuclease	arCOG05724
AD	cHG_038	PredictedRestrictionEndonuclease	arCOG06431
AE	cHG_015	HNH endonuclease	arCOG07787
AF	cHG_019	Endonuclease	arCOG02782
AG	cHG_020	Endonuclease	arCOG02781
AH	cHG_004	HNH endonuclease	arCOG09398
AI	cHG_037	HNH_nuclease	arCOG05223
AJ	cHG_039	HNH nuclease	arCOG03898
AK	cHG_041	HNH nuclease	arCOG08099
AL	cHG_046	MBF1	arCOG01863
AM	cHG_028	CBS domain	arCOG00608
AN	cHG_005	MarR	arCOG03182
AO	cHG_030	ParB-like nuclease	arCOG01875
AP	cHG_016	GVPC	arCOG06392
AQ	cHG_002	ASCH domain RNA binding	arCOG01734
AR	cHG_017	Uncharacterized	arCOG10082
AS	cHG_026	Uncharacterized	arCOG13171
AT	cHG_032	Uncharacterized	arCOG08946
AU	cHG_043	Uncharacterized	arCOG08856
AV	cHG_047	Uncharacterized	arCOG04588

[$]: A listing of associated Gene Ontology terms and gene family descriptions is available in supplementary Table S2

[$$]: T_I and T_II denote type I and type II restriction enzyme, respectively. M, R, S denote the methylase, restriction endonuclease, and specificity subunits, respectively.

Table 2. Important traits of cHGs with four or more ORFs.

Alpha (numeric) cHG	No. of taxa	No. of transfers[a]	Function[b]	Predicted Recognition sites[c]	Frequency[e]
I (001)	16	9	T_II_M	GAAGGC	31%
				GGRCA	31%
J (003)	38	21	T_II_M	CANCATC	53%
				TAGGAG	21%
AH (004)	12	4	HNH_endonuclease	GGCGCC	89%
				GATC	11%
F (006)	61	44	T_I_S	GGAYNNNNNNTGG	24%
				CAGNNNNNNTGCT	16%
R (008)	14	0	T_III_R	NA[d]	100%
AA (010)	55	15	RNA_methylase	ATTAAT	33%
K (011)	137	97	T_II_M	GCAAGG	49%
				GKAAYG	28%
AC (012)	8	5	Restriction Endonuclease	GCGAA	29%
				CAACNNNNNTC	29%
				CTGGAG	29%
T (014)	130	93	Adenine_DNA_methylase	GCAGG	45%
				AAGCTT	32%
AE (015)	21	13	HNH_endonuclease	GGCGCC	70%
				YSCNS	15%
AP (016)	12	6	GVPC	CANCATC	83%
C (018)	7	4	T_I_R	AACNNNNNNGTGC	73%
				CTANNNNNNRTTC	27%
AF (019)	4	3	Endonuclease	NA[d]	100%
A (021)	88	58	T_I_M	GGAYNNNNNNTGG	37%
				GTCANNNNNNRTCA	12%
				CTCGAG	9%
U (022)	290	120	DNA_methylase	CTAG	59%
				CATTC	14%
				CCCGGG	7%
O (023)	37	28	T_II_R	NA[d]	100%
B (024)	16	8	T_I_M	GAGNNNNNNVTGAC	75%
				GACNNNNNNRTAC	19%
G (025)	4	2	T_I_S	GAGNNNNRTAA	75%
				GAGNNNNNTAC	25%
V (027)	5	1	DNA_methylase	CATTC	100%
AO (030)	4	2	ParB-like_nuclease	GATC	75%
				CTAG	25%
W (031)	153	70	dam_methylase	GATC	70%
				AB / SAAM	22%

AT (032)	116	60	Uncharacterized	GCAAGG	43%
				GKAAYG	26%
				GGTTAG	14%
L (033)	66	38	T_II_M-033	CAARCA	40%
				CTGAAG	36%
D (034)	16	11	T_I_R-034	GCANNNNNRTTA	69%
				GGCANNNNNNTTC	19%
X (035)	19	9	probable_RMS_M	GGGAC	83%
H (036)	38	24	probable_T_II_M	CCWGG	42%
				CCSGG	18%
				GTAC	16%
AI (037)	6	4	HNH_nuclease	NA[d]	100%
AJ (039)	5	4	HNH_nuclease	GGCGCC	100%
AK (041)	6	4	HNH_nuclease	NA[d]	100%
Q (042)	21	8	Adenine_DNA_methylase	RGTAAT	71%
			probable_T_III_M	NA[d]	19%
Y (044)	179	110	dcm_methylase	CGGCCG	24%
				GTCGAC	13%
				ACGT	11%
E (045)	58	42	T_I_R	CCCNNNNNRTTGY	63%
				GCANNNNNRTTA	28%
Z (048)	54	35	Adenine_DNA_methylase	CCRGAG	36%
				GTMKAC	30%

[a] Number of estimated horizontal gene transfer events

[b] T_I and T_II denote type I and type II restriction enzyme, respectively. M, R, S denote the methylase, restriction endonuclease, and specificity subunits, respectively.

[c] Top predicted recognition sites

[d] No predicted recognition site

[e] Frequency of predictions within the cHG

381

382

4. Discussion

A striking result of our study is the irregular distribution of the RM system gene candidates throughout not just the haloarchaeal class, but also within it's orders, genera, species, and even communities and populations. The patchy distribution is almost certainly the result of frequent HGT and gene loss. RM system genes are well known for their susceptibility to HGT and loss, and their presence almost never define a clade or an environmental source (e.g., [36,76]). Frequent acquisition of RM system genes through HGT is illustrated by their sporadic distribution. For example, *Halorubrum* genomes encode many candidate RM system cHGs that are absent from the remainder of the Halobacteria (e.g., cHG M, R, S, AC, AG, AM). Only one of these (cHG R) is found in more than 3 genomes, a Type III restriction protein found in 14 of 57 *Halorubrum* genomes. Gene loss undoubtedly contributed to the sparse cHGs distribution; however, without invoking frequent gene transfer, many independent and parallel gene losses need to be postulated. We also observed that a number haloarchaeal species possess multiple Type I subunit genes, allowing for functional substitution of the different subunits in the RM system. The existence of multiple Type I subunits has also been observed in *Helicobacter pylori*, in which 4 different SSU loci are used by the organism's Type I system to target different recognition sequences; these SSUs can even exchange TRDs, resulting in variation in the methylome of *H. pylori* [77–79]. In our results, however, we observed multiple MTase and REase subunits alongside a single SSU, suggesting the functional substitution of the subunits in these haloarchaeal organisms does not result in variation in detected recognition sequences.

It seems counterintuitive that RM systems are not more conserved as cellular countermeasures against commonly occurring viruses. It may be that cells do not require extensive protection via RM systems, because they use multiple defensive systems some of which might be more effective. For example, another well-known defense against viruses is the CRISPR-Cas system [80]. CRISPR recognizes short (~40bp) regions of invading DNA that the host has been exposed to previously and degrades it. While it can be very useful against virus infection, our prior work indicated that CRISPR-Cas was also sporadically distributed within communities of closely related haloarchaeal species [81] indicating they are not required for surviving virus infection.

Both the RM and CRISPR-Cas systems are only important countermeasures after external fortifications have failed to prevent a virus from infiltrating, and therefore their limited distributions also indicate that the cell's primary defense would be in preventing virus infection altogether, which is accomplished by different mechanisms. By altering surfaces via glycosylation cells can avoid virus predation prior to infection. In *Haloferax* species there are two pathways which control glycosylation of external features. One is relatively conserved and could have functions other than virus avoidence, while the other is highly variable and shows hallmarks of having genes mobilized by horizontal transfer [82]. At least one halovirus has been found to require glycosylation by its host in order to infect properly [83]. Comparison of genomes and metagenomes from hypersaline envirionments showed widespread evidence for distinct "genomic" islands in closely related halophiles [84] that contain a unique mixture of LPS and other genes that contribute to altering the cell's surface structure and virus docking opportunities. Thus selective pressure on post infection, cytosolic and nucleic acids-based virus defenses is eased, allowing them to be lost randomly in populations.

A major consideration in understanding RM system diversity is that viruses, or other infiltrating selfish genetic elements, might gain access to the host's methylation after a successful infection that was not stopped by the restriction system. Indeed, haloviruses are known to encode DNA methyltransferases in their genomes (e.g., see [85]). In this case, RM systems having a limited within population distribution would then be an effective defense for that part of the population possessing a different RM system. Under this scenario, a large and diverse pool of mobilized RM system genes could offer a stronger defense for the population as a whole. A single successful infection would no longer endanger the entire group of potential hosts.

Group selection may be invoked to explain the within population diversity of RM systems; a sparse distribution of RM systems may provide a potential benefit to the population as a whole, because a virus cannot easily infect all members of the population. However, often gene level selection is a more appropriate alternative to group selection [86,87]. Under a gene centered explanation, RM systems are considered as selfish addiction cassettes that may be of little benefit to its carrier. While RM systems may be difficult to delete as a whole, stepwise deletion, that begins with inactivation of the REase activity can lead to their loss from a lineage. Their long-term survival thus may be a balance of gain through gene transfer, persistence through addiction, and gene loss. This gene centered explanation is supported by a study from [36], which examined the distribution of MTase genes in ~1000 bacterial genomes. They observed, similar to our results in the Halobacteria, that MTases associated with RM systems are poorly conserved, whereas orphan MTases share conservation patterns similar to average genes. They also demonstrated that many RM-associated and orphan MTases are horizontally acquired, and that a number of orphan MTases in bacterial genomes neighbor degraded REase genes, suggesting that they are the product of degraded RM systems that have lost functional REases [36]. Similarly, Kong et al. [76] studying genome content variation in *Neisseria meningitidis* found an irregular distribution of RM systems, suggesting that these systems do not form an effective barrier to homologous recombination within the species. Kong et al. also observed that the RM systems themselves had been frequently transferred within the species. We conclude that RM genes in bacteria as well as archaea appear to undergo significant horizontal transfer and are not well-conserved. Only when these genes pick up additional functions, do parts of these systems persist for longer periods of time, as exemplified in the distribution of orphan MTases. However, the transition from RM system MTase to orphan MTase is an infrequent event. A study of 43 pan-genomes by Oliveira et al. [88] suggests that orphan MTases occur more frequently from transfer via large mobile genetic elements (MGEs) such as plasmids and phages rather than arise *de novo* from RM degradation. The distribution of orphan methylase cHG U and W, and their likely target motifs, CTAG and GATC, respectively suggests different biological functions for these two methylases. Similar to other bacterial and archaeal genomes [89], the CTAG motif, the likely target for methylases in cHG U, is underrepresented in all haloarchaeal genomes (see table S3). The low frequency of occurrence, only about once per 4000 nucleotides, suggests that this motif and the cognate orphan methylase are not significantly involved in facilitating mismatch repair. The underrepresented CTAG motif was found to be less underrepresented near rRNA genes [89] and on plasmids; the CTAG motif also is a known target sequence for some IS elements [90]; and it may be involved in repressor binding, where the CTAG motif was found to be associated with kinks in the DNA when bound to the repressor [91,92] Interestingly, CTAG and GATC motifs are absent, or underrepresented in several haloarchaeal viruses [85,93,94]. However, at present the reasons for the underrepresentation of the motif in chromosomal DNA, and the role that the methylation of this motif may play remain open questions.

5. Conclusions

RM systems have a sporadic distribution in haloarchaea, even within species and populations. In contrast, orphan methylases are more persistent in lineages, and the targeted motifs are under selection for lower (in case of CTAG) or higher (in case of GATC) than expected frequency. In case of the GATC motif, the cognate orphan MTase was found only in genomes where this motif occurs with high frequency.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1,

Figure S1. Workflow of RMS-candidate gene search strategy.

Figure S2. Plot of alignment distance as a function of presence-absence distance.

Figure S3. PCoA plot of the distances between the RMS presence-absence profiles of the 217 analyzed Halobacterial genomes.

Figure S4. Maximum-likelihood phylogeny of cHG presence-absence matrix.

Figure S5. Bootstrap support values of the presence-absence phylogeny mapped onto the ribosomal protein reference tree.

Table S1. Basic statistics for Halobacteriaceae complete and draft genomes

Table S2. Gene Ontology (GO) terms for each collapsed homologous group

Table S3. Distribution of orphan methylases cHGs U and W and frequency of their putative recognition motifs in completely sequenced halobacterial chromosomes.

Author Contributions: Conceptualization, Thane R. Papke and Johann Peter Gogarten; Data curation, Matthew S. Fullmer, Matthew Ouellette and Artemis S. Louyakis; Formal analysis, Matthew S. Fullmer, Matthew Ouellette and Artemis S. Louyakis; Funding acquisition, Thane R. Papke and Johann Peter Gogarten; Investigation, Matthew S. Fullmer and Matthew Ouellette; Methodology, Matthew S. Fullmer, Artemis S. Louyakis and Johann Peter Gogarten; Project administration, Thane R. Papke and Johann Peter Gogarten; Software, Matthew S. Fullmer and Artemis S. Louyakis; Supervision, Thane R. Papke and Johann Peter Gogarten; Validation, Matthew S. Fullmer; Visualization, Matthew S. Fullmer and Artemis S. Louyakis; Writing – original draft, Matthew S. Fullmer and Johann Peter Gogarten; Writing – review & editing, Matthew Ouellette, Artemis S. Louyakis, Thane R. Papke and Johann Peter Gogarten.

Funding: This work was supported through grants from the Binational Science Foundation (BSF 2013061); the National Science Foundation (NSF/MCB 1716046) within the BSF-NSF joint research program; and NASA exobiology (NNX15AM09G, and 80NSSC18K1533).

Acknowledgments: The Computational Biology Core, Institute for Systems Genomics, University of Connecticut provided computational resources.

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

References

1. Korlach, J.; Turner, S. W. Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.* **2012**, *22*, 251–261, doi:10.1016/j.sbi.2012.04.002.
2. Bheemanaik, S.; Reddy, Y. V. R.; Rao, D. N. Structure, function and mechanism of exocyclic DNA methyltransferases. *Biochem. J.* **2006**, *399*, 177–190, doi:10.1042/BJ20060854.
3. Malone, T.; Blumenthal, R. M.; Cheng, X. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.* **1995**, *253*, 618–632, doi:10.1006/jmbi.1995.0577.
4. Bujnicki, J. M.; Radlinska, M. Molecular evolution of DNA-(cytosine-N4) methyltransferases: evidence for their polyphyletic origin. *Nucleic Acids Res.* **1999**, *27*, 4501–9.
5. Bujnicki, J. M. Sequence permutations in the molecular evolution of DNA methyltransferases. *BMC Evol. Biol.* **2002**, *2*, 3.
6. Tock, M. R.; Dryden, D. T. The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* **2005**, *8*, 466–472, doi:10.1016/j.mib.2005.06.003.
7. Vasu, K.; Nagaraja, V. Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 53–72, doi:10.1128/MMBR.00044-12.
8. Pleška, M.; Qian, L.; Okura, R.; Bergmiller, T.; Wakamoto, Y.; Kussell, E.; Guet, C. C. Bacterial autoimmunity due to a restriction-modification system. *Curr. Biol. CB* **2016**, *26*, 404–409, doi:10.1016/j.cub.2015.12.041.
9. Ohno, S.; Handa, N.; Watanabe-Matsui, M.; Takahashi, N.; Kobayashi, I. Maintenance forced by a restriction-modification system can be modulated by a region in its modification enzyme not essential for methyltransferase activity. *J. Bacteriol.* **2008**, *190*, 2039–2049, doi:10.1128/JB.01319-07.
10. Kobayashi, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **2001**, *29*, 3742–56.
11. Budroni, S.; Siena, E.; Hotopp, J. C. D.; Seib, K. L.; Serruto, D.; Nofroni, C.; Comanducci, M.; Riley, D. R.; Daugherty, S. C.; Angiuoli, S. V.; Covacci, A.; Pizza, M.; Rappuoli, R.; Moxon, E. R.; Tettelin, H.; Medini, D. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl. Acad. Sci.* **2011**, *108*, 4494–4499, doi:10.1073/pnas.1019751108.
12. Erwin, A. L.; Sandstedt, S. A.; Bonthuis, P. J.; Geelhood, J. L.; Nelson, K. L.; Unrath, W. C. T.; Diggle, M. A.; Theodore, M. J.; Pleatman, C. R.; Mothershed, E. A.; Sacchi, C. T.; Mayer, L. W.; Gilsdorf, J. R.; Smith, A. L. Analysis of Genetic Relatedness of Haemophilus influenzae Isolates by Multilocus Sequence Typing. *J. Bacteriol.* **2008**, *190*, 1473–1483, doi:10.1128/JB.01207-07.
13. Roer, L.; Aarestrup, F. M.; Hasman, H. The EcoKI type I restriction-modification system in escherichia coli affects but is not an absolute barrier for conjugation. *J. Bacteriol.* **2015**, *197*, 337–342, doi:10.1128/JB.02418-14.
14. McKane, M.; Milkman, R. Transduction, restriction and recombination patterns in Escherichia coli. *Genetics* **1995**, *139*, 35–43.
15. Chang, S.; Cohen, S. N. In vivo site-specific genetic recombination promoted by the EcoRI restriction endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74*, 4811–5.
16. Lin, E. A.; Zhang, X.-S.; Levine, S. M.; Gill, S. R.; Falush, D.; Blaser, M. J. Natural transformation of helicobacter pylori involves the integration of short dna fragments interrupted by gaps of variable size. *PLoS Pathog.* **2009**, *5*, e1000337, doi:10.1371/journal.ppat.1000337.
17. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res.* **1964**, *1*, 2–9.
18. Ershova, A. S.; Rusinov, I. S.; Spirin, S. A.; Karyagina, A. S.; Alexeevski, A. V. Role of restriction-modification systems in prokaryotic evolution and ecology. *Biochem.* **2015**, *80*, 1373–1386, doi:10.1134/S0006297915100193.
19. Roberts, R. J.; Belfort, M.; Bestor, T.; Bhagwat, A. S.; Bickle, T. A.; Bitinaite, J.; Blumenthal, R. M.; Degtyarev, S. K.; Dryden, D. T. F.; Dybvig, K.; Firman, K.; Gromova, E. S.; Gumport, R. I.; Halford, S. E.; Hattman, S.; Heitman, J.; Hornby, D. P.; Janulaitis, A.; Jeltsch, A.; Josephsen, J.; Kiss, A.; Klaenhammer, T. R.; Kobayashi, I.; Kong, H.; Krüger, D. H.; Lacks, S.; Marinus, M. G.; Miyahara, M.; Morgan, R. D.; Murray, N. E.; Nagaraja, V.; Piekarowicz, A.; Pingoud, A.; Raleigh, E.; Rao, D. N.; Reich, N.; Repin, V. E.; Selker, E. U.; Shaw, P.-C.; Stein, D. C.; Stoddard, B. L.; Szybalski, W.; Trautner, T. A.; Van Etten, J. L.; Vitor, J. M. B.; Wilson, G. G.; Xu, S. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **2003**, *31*, 1805–12.

20. Loenen, W. A. M.; Dryden, D. T. F.; Raleigh, E. A.; Wilson, G. G. Type I restriction enzymes and their relatives. *Nucleic Acids Res.* **2014**, *42*, 20–44, doi:10.1093/nar/gkt847.
21. Liu, Y.-P.; Tang, Q.; Zhang, J.-Z.; Tian, L.-F.; Gao, P.; Yan, X.-X. Structural basis underlying complex assembly and conformational transition of the type I R-M system. *Proc. Natl. Acad. Sci.* **2017**, *114*, 11151–11156, doi:10.1073/pnas.1711754114.
22. Pingoud, A.; Wilson, G. G.; Wende, W. Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res.* **2014**, *42*, 7489–7527, doi:10.1093/nar/gku447.
23. Morgan, R. D.; Bhatia, T. K.; Lovasco, L.; Davis, T. B. MmeI: a minimal Type II restriction-modification system that only modifies one DNA strand for host protection. *Nucleic Acids Res.* **2008**, *36*, 6558–6570, doi:10.1093/nar/gkn711.
24. Rao, D. N.; Dryden, D. T. F.; Bheemanaik, S. Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res.* **2014**, *42*, 45–55, doi:10.1093/nar/gkt616.
25. Czapinska, H.; Kowalska, M.; Zagorskaitė, E.; Manakova, E.; Slyvka, A.; Xu, S.; Siksnys, V.; Sasnauskas, G.; Bochtler, M. Activity and structure of EcoKMcrA. *Nucleic Acids Res.* **2018**, *46*, 9829–9841, doi:10.1093/nar/gky731.
26. Adhikari, S.; Curtis, P. D. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* **2016**, *40*, 575–591, doi:10.1093/femsre/fuw023.
27. Messer, W.; Bellekes, U.; Lother, H. Effect of dam methylation on the activity of the E. coli replication origin, oriC. *EMBO J.* **1985**, *4*, 1327–32.
28. Kang, S.; Lee, H.; Han, J. S.; Hwang, D. S. Interaction of SeqA and Dam methylase on the hemimethylated origin of Escherichia coli chromosomal DNA replication. *J. Biol. Chem.* **1999**, *274*, 11463–8.
29. Sanchez-Romero, M. A.; Busby, S. J. W.; Dyer, N. P.; Ott, S.; Millard, A. D.; Grainger, D. C. Dynamic distribution of SeqA protein across the chromosome of Escherichia coli K-12. *MBio* **2010**, *1*, doi:10.1128/mBio.00012-10.
30. Welsh, K. M.; Lu, A. L.; Clark, S.; Modrich, P. Isolation and characterization of the Escherichia coli mutH gene product. *J. Biol. Chem.* **1987**, *262*, 15624–9.
31. Au, K. G.; Welsh, K.; Modrich, P. Initiation of methyl-directed mismatch repair. *J. Biol. Chem.* **1992**, *267*, 12142–8.
32. Putnam, C. D. Evolution of the methyl directed mismatch repair system in Escherichia coli. *DNA Repair (Amst).* **2016**, *38*, 32–41, doi:10.1016/j.dnarep.2015.11.016.
33. Zweiger, G.; Marczynski, G.; Shapiro, L. A Caulobacter DNA Methyltransferase that functions only in the predivisional cell. *J. Mol. Biol.* **1994**, *235*, 472–485, doi:10.1006/jmbi.1994.1007.
34. Domian, I. J.; Reisenauer, A.; Shapiro, L. Feedback control of a master bacterial cell-cycle regulator. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 6648–53.
35. Gonzalez, D.; Kozdon, J. B.; McAdams, H. H.; Shapiro, L.; Collier, J. The functions of DNA methylation by CcrM in Caulobacter crescentus: a global approach. *Nucleic Acids Res.* **2014**, *42*, 3720–3735, doi:10.1093/nar/gkt1352.
36. Seshasayee, A. S. N.; Singh, P.; Krishna, S. Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.* **2012**, *40*, 7066–7073, doi:10.1093/nar/gks390.
37. Blow, M. J.; Clark, T. A.; Daum, C. G.; Deutschbauer, A. M.; Fomenkov, A.; Fries, R.; Froula, J.; Kang, D. D.; Malmstrom, R. R.; Morgan, R. D.; Posfai, J.; Singh, K.; Visel, A.; Wetmore, K.; Zhao, Z.; Rubin, E. M.; Korlach, J.; Pennacchio, L. A.; Roberts, R. J. The epigenomic landscape of prokaryotes. *PLOS Genet.* **2016**, *12*, e1005854, doi:10.1371/journal.pgen.1005854.
38. Nölling, J.; de Vos, W. M. Identification of the CTAG-recognizing restriction-modification systems MthZI and MthFI from Methanobacterium thermoformicicum and characterization of the plasmid-encoded mthZIM gene. *Nucleic Acids Res.* **1992**, *20*, 5047–52.
39. Grogan, D. W. Cytosine methylation by the SuaI restriction-modification system: implications for genetic fidelity in a hyperthermophilic archaeon. *J. Bacteriol.* **2003**, *185*, 4657–61.
40. Ishikawa, K.; Watanabe, M.; Kuroita, T.; Uchiyama, I.; Bujnicki, J. M.; Kawakami, B.; Tanokura, M.; Kobayashi, I. Discovery of a novel restriction endonuclease by genome comparison and application of a wheat-germ-based cell-free translation assay: PabI (5′-GTA/C) from the hyperthermophilic archaeon Pyrococcus abyssi. *Nucleic Acids Res.* **2005**, *33*, e112–e112, doi:10.1093/nar/gni113.

41. Watanabe, M.; Yuzawa, H.; Handa, N.; Kobayashi, I. Hyperthermophilic DNA methyltransferase M.PabI from the archaeon Pyrococcus abyssi. *Appl. Environ. Microbiol.* **2006**, *72*, 5367–5375, doi:10.1128/AEM.00433-06.
42. Couturier, M.; Lindås, A.-C. The DNA methylome of the hyperthermoacidophilic crenarchaeon Sulfolobus acidocaldarius. *Front. Microbiol.* **2018**, *9*, 137, doi:10.3389/fmicb.2018.00137.
43. Chimileski, S.; Dolas, K.; Naor, A.; Gophna, U.; Papke, R. T. Extracellular DNA metabolism in Haloferax volcanii. *Front. Microbiol.* **2014**, *5*, 57, doi:10.3389/fmicb.2014.00057.
44. Zerulla, K.; Chimileski, S.; Näther, D.; Gophna, U.; Papke, R. T.; Soppa, J. DNA as a Phosphate storage polymer and the alternative advantages of polyploidy for growth or survival. *PLoS One* **2014**, *9*, e94819, doi:10.1371/journal.pone.0094819.
45. Ouellette, M.; Gogarten, J.; Lajoie, J.; Makkay, A.; Papke, R. Characterizing the DNA methyltransferases of Haloferax volcanii via bioinformatics, gene deletion, and SMRT sequencing. *Genes (Basel).* **2018**, *9*, 129, doi:10.3390/genes9030129.
46. Ouellette, M.; Jackson, L.; Chimileski, S.; Papke, R. T. Genome-wide DNA methylation analysis of Haloferax volcanii H26 and identification of DNA methyltransferase related PD-(D/E)XK nuclease family protein HVO_A0006. *Front. Microbiol.* **2015**, *6*, 251, doi:10.3389/fmicb.2015.00251.
47. Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **2015**, gr-186072.
48. Roberts, R. J.; Macelis, D. REBASE—restriction enzymes and methylases. *Nucleic Acids Res.* **2001**, *29*, 268–269, doi:10.1093/nar/29.1.268.
49. Roberts, R. J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **2015**, *43*, D298–D299, doi:10.1093/nar/gku1046.
50. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461, doi:10.1093/bioinformatics/btq461.
51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **2004**, *32*, 1792–1797.
52. Makarova, K. S.; Wolf, Y. I.; Koonin, E. V. Archaeal clusters of orthologous genes (ARCOGS): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel, Switzerland)* **2015**, *5*, 818–840, doi:10.3390/life5010818.
53. 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.
54. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal* **2006**, *Complex Sy*, 1695.
55. Caspi, R.; Altman, T.; Dale, J. M.; Dreher, K.; Fulcher, C. A.; Gilham, F.; Kaipa, P.; Karthikeyan, A. S.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Paley, S.; Popescu, L.; Pujar, A.; Shearer, A. G.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2010**, *38*, D473–D479, doi:10.1093/nar/gkp875.
56. Hoang, D. T.; Chernomor, O.; von Haeseler, A.; Minh, B. Q.; Le, S. V. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*, doi:10.1093/molbev/msx281.
57. 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.
58. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313, doi:10.1093/bioinformatics/btu033.
59. Bansal, M. S.; Wu, Y.-C.; Alm, E. J.; Kellis, M. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* **2015**, *31*, 1211–1218, doi:10.1093/bioinformatics/btu806.
60. Bansal, M. S.; Alm, E. J.; Kellis, M. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* **2012**, *28*, i283–i291, doi:10.1093/bioinformatics/bts225.
61. Yu, G.; Smith, D. K.; Zhu, H.; Guan, Y.; Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **2017**, *8*, 28–36, doi:10.1111/2041-210X.12628.

62. Yu, G.; Lam, T. T.-Y.; Zhu, H.; Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Mol. Biol. Evol.* **2018**, *35*, 3041–3043, doi:10.1093/molbev/msy194.
63. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **2003**, *14*, 927–930, doi:10.1111/j.1654-1103.2003.tb02228.x.
64. Wickham, H. *ggplot2: elegant graphics for data analysis*; Springer, 2016; ISBN 3319242776.
65. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **2011**, *27*, 592–3, doi:10.1093/bioinformatics/btq706.
66. Salichos, L.; Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **2013**, *497*, 327–331, doi:10.1038/nature12130.
67. Drost, H. Philentropy: Information Theory and Distance Quantification with R. *J. Open Source Softw.* **2018**, *3*.
68. Gogarten, J. P. Perl script to measure frequency and distribution of CTAG and GATC motifs in DNA Available online: https://github.com/Gogarten-Lab/CTAG-GATC-frequencies (accessed on Jan 12, 2019).
69. Conesa, A.; Götz, S.; García-Gómez, J. M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676.
70. Consortium, U. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2016**, *45*, D158–D169.
71. Soucy, S. M.; Fullmer, M. S.; Papke, R. T.; Gogarten, J. P. Inteins as indicators of gene flow in the halobacteria. *Front. Microbiol.* **2014**, *5*, 299, doi:10.3389/fmicb.2014.00299.
72. Gupta, R. S.; Naushad, S.; Fabros, R.; Adeolu, M. A phylogenomic reappraisal of family-level divisions within the class Halobacteria: proposal to divide the order Halobacteriales into the families Halobacteriaceae, Haloarculaceae fam. nov., and Halococcaceae fam. nov., and the order Haloferacales into th. *Antonie Van Leeuwenhoek* **2016**, *109*, 565–587, doi:10.1007/s10482-016-0660-2.
73. Gupta, R. S.; Naushad, S.; Baker, S. Phylogenomic analyses and molecular signatures for the class Halobacteria and its two major clades: a proposal for division of the class Halobacteria into an emended order Halobacteriales and two new orders, Haloferacales ord. nov. and Natrialbales ord. n. *Int. J. Syst. Evol. Microbiol.* **2015**, *65*, 1050–69, doi:10.1099/ijs.0.070136-0.
74. Ohno, S.; Handa, N.; Watanabe-Matsui, M.; Takahashi, N.; Kobayashi, I. Maintenance forced by a restriction-modification system can be modulated by a region in its modification enzyme not essential for methyltransferase activity. *J. Bacteriol.* **2008**, *190*, 2039–2049, doi:10.1128/JB.01319-07.
75. Roberts, R. J.; Belfort, M.; Bestor, T.; Bhagwat, A. S.; Bickle, T. A.; Bitinaite, J.; Blumenthal, R. M.; Degtyarev, S. K.; Dryden, D. T. F.; Dybvig, K.; Firman, K.; Gromova, E. S.; Gumport, R. I.; Halford, S. E.; Hattman, S.; Heitman, J.; Hornby, D. P.; Janulaitis, A.; Jeltsch, A.; Josephsen, J.; Kiss, A.; Klaenhammer, T. R.; Kobayashi, I.; Kong, H.; Krüger, D. H.; Lacks, S.; Marinus, M. G.; Miyahara, M.; Morgan, R. D.; Murray, N. E.; Nagaraja, V.; Piekarowicz, A.; Pingoud, A.; Raleigh, E.; Rao, D. N.; Reich, N.; Repin, V. E.; Selker, E. U.; Shaw, P.-C.; Stein, D. C.; Stoddard, B. L.; Szybalski, W.; Trautner, T. A.; Etten, V.; L, J.; Vitor, J. M. B.; Wilson, G. G.; Xu, S. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **2003**, *31*, 1805–1812, doi:10.1093/nar/gkg274.
76. 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*, 1611–27, doi:10.1093/gbe/evt116.
77. Furuta, Y.; Namba-Fukuyo, H.; Shibata, T. F.; Nishiyama, T.; Shigenobu, S.; Suzuki, Y.; Sugano, S.; Hasebe, M.; Kobayashi, I. Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLoS Genet.* **2014**, *10*, e1004272, doi:10.1371/journal.pgen.1004272.
78. Furuta, Y.; Kobayashi, I. Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics-driven adaptive evolution. *Mob. Genet. Elements* **2012**, *2*, 292–296, doi:10.4161/mge.23371.
79. Furuta, Y.; Kawai, M.; Uchiyama, I.; Kobayashi, I. Domain movement within a gene: a novel evolutionary mechanism for protein diversification. *PLoS One* **2011**, *6*, e18819, doi:10.1371/journal.pone.0018819.
80. Gophna, U.; Brodt, A. CRISPR/Cas systems in archaea. *Mob. Genet. Elements* **2012**, *2*, 63–64, doi:10.4161/mge.19907.
81. Fullmer, M. S.; Soucy, S. M.; Swithers, K. S.; Makkay, A. M.; Wheeler, R.; Ventosa, A.; Gogarten, J. P.; Papke, R. T. Population and genomic analysis of the genus Halorubrum. *Front. Microbiol.* **2014**, *5*, 140, doi:10.3389/fmicb.2014.00140.

82. Shalev, Y.; Soucy, S.; Papke, R.; Gogarten, J.; Eichler, J.; Gophna, U. Comparative analysis of surface layer glycoproteins and genes involved in protein glycosylation in the genus Haloferax. *Genes (Basel).* **2018**, *9*, 172, doi:10.3390/genes9030172.
83. Kandiba, L.; Aitio, O.; Helin, J.; Guan, Z.; Permi, P.; Bamford, D. H.; Eichler, J.; Roine, E. Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. *Mol. Microbiol.* **2012**, *84*, 578–593, doi:10.1111/j.1365-2958.2012.08045.x.
84. Rodriguez-Valera, F.; Martin-Cuadrado, A.-B.; Rodriguez-Brito, B.; Pašić, L.; Thingstad, T. F.; Rohwer, F.; Mira, A. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **2009**, *7*, 828–836, doi:10.1038/nrmicro2235.
85. Tang, S.-L.; Nuttall, S.; Ngui, K.; Fisher, C.; Lopez, P.; Dyall-Smith, M. HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. *Mol. Microbiol.* **2002**, *44*, 283–96.
86. Olendzenski, L.; Gogarten, J. P. Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. *Ann N Y Acad Sci* **2009**, *1178*, 137–145, doi:NYAS4998 [pii]10.1111/j.1749-6632.2009.04998.x.
87. Naor, A.; Altman-Price, N.; Soucy, S. M.; Green, A. G.; Mitiagin, Y.; Turgeman-Grott, 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*, E4654-61, doi:10.1073/pnas.1606416113.
88. Oliveira, P. H.; Touchon, M.; Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **2014**, *42*, 10618–31, doi:10.1093/nar/gku734.
89. Karlin, S.; Mrázek, J.; Campbell, A. M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **1997**, *179*, 3899–913.
90. Fournier, P.; Paulus, F.; Otten, L. IS870 requires a 5′-CTAG-3′ target sequence to generate the stop codon for its large ORF1. *J. Bacteriol.* **1993**, *175*, 3151–60.
91. Otwinowski, Z.; Schevitz, R. W.; Zhang, R.-G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. Crystal structure of trp represser/operator complex at atomic resolution. *Nature* **1988**, *335*, 321–329, doi:10.1038/335321a0.
92. Burge, C.; Campbell, A. M.; Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 1358–62.
93. Bath, C.; Cukalac, T.; Porter, K.; Dyall-Smith, M. L. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. *Virology* **2006**, *350*, 228–39, doi:10.1016/j.virol.2006.02.005.
94. Porter, K.; Tang, S.-L.; Chen, C.-P.; Chiang, P.-W.; Hong, M.-J.; Dyall-Smith, M. PH1: an archaeovirus of Haloarcula hispanica related to SH1 and HHIV-2. *Archaea* **2013**, *2013*, 456318, doi:10.1155/2013/456318.