Title: Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation

Authors: Israela Turgeman-Grott[1], Shirley Joseph[1], Sam Marton[1], Kim Eizenshtein[1], Adit Naor[1†], Shannon Soucy[2], Aris-Edda Stachler[3], Yarden Shalev[1], Mor Zarkor[1], Leah Reshef[1], Neta Altman-Price [1], Anita Marchfelder[3], and Uri Gophna[1*].

Affiliations:

[1] Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978-01, Israel.

[2] Department of Biological Sciences, Dartmouth College, 78 College Street, Hanover, New Hampshire, 03766, USA.

[3] Department of Biology II, Ulm University, 89069 Ulm, Germany.

[†]Present address: Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305.

*Correspondence urigo@tauex.tau.ac.il.

Summary

CRISPR-Cas systems provide prokaryotes with sequence-specific immunity against viruses and plasmids, based on DNA acquired from these invaders, known as spacers. Surprisingly, many archaea possess spacers that match chromosomal genes of related species, including those encoding core housekeeping genes. By sequencing genomes of environmental archaea isolated from a single site, we demonstrate that inter-species spacers are common. We show experimentally by mating *Haloferax volcanii* and *Haloferax mediterranei*, that spacers are indeed acquired chromosome-wide, although a preference for integrated mobile elements and nearby regions of the chromosome exists. Inter-species mating induces increased spacer acquisition and may result in interactions between the acquisition machinery of the two species. Surprisingly, many of the spacers acquired following inter-species mating target self-replicons along with those originating from the mating partner, indicating that the acquisition machinery cannot distinguish self from non-self under these conditions. Engineering the chromosome of one species to be targeted by the other's CRISPR-Cas reduces gene exchange between them substantially. Thus, spacers acquired during inter-species mating could limit future gene transfer, resulting in a role for CRISPR-Cas systems in microbial speciation.

Keywords

CRISPR ; archaea; horizontal gene transfer; lateral gene transfer; adaptation; acquisition; speciation; mobile genetic elements

Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins) systems provide acquired heritable immunity to bacteria and archaea against invasion by selfish DNA elements. CRISPR loci are composed of partially palindromic repeats interspersed by short unique DNA spacers and multiple *cas* genes that encode proteins involved in the immune response. These systems can acquire DNA fragments from foreign selfish elements and integrate them as spacers into the CRISPR arrays [1,2]. Subsequently, One or several spacer arrays in a prokaryotic cell can be transcribed and processed into small CRISPR RNA (crRNA) molecules, which then together with the Cascade (CRISPR-associated complex for antiviral defense) complex hybridize with nucleotide sequences in the invader and direct its degradation by Cas endonucleases [1,3].

CRISPR-Cas loci are considered to be primarily anti-viral defense systems, and in some lineages whose viruses have been studied, such as the Sulfolobales, that is reflected by the vast majority of spacers matching viral sequence [4,5]. Strikingly, however, the vast majority of CRISPR spacers in bacteria and archaea that have database hits match integrated proviruses rather than lytic bacteriophages or archaeal viruses [7]. This can be attributed to the under-sampling of virus sequence space, but may also indicate that excision from host chromosomes represents a preferred opportunity for spacer incorporation as much as viral DNA injection into a host cell [8].

We previously observed that multiple archaea belonging to diverse clades have CRISPR spacers that match chromosomal housekeeping genes of related species, rather than selfish elements [9]. This raises the question of how such spacers were acquired and whether they can affect gene exchange dynamics between species, when present. Notably, halophilic archaea can undergo a mating by cell fusion process involving cytoplasmic bridges [10], which can efficiently occur between cells from different species [11]. Similar cytoplasmic bridges between cells have also been observed in multiple other archaeal lineages, such as members of Sulfolobales [12], Thermococcales [13], and most recently, nanoarchaea and Thermoplasmatales [14]. Here we test the hypothesis that spacers can be acquired naturally during inter-species

mating from partner chromosomes. We use genome sequences of 15 haloarchaea isolated from the same coastal site to show that inter-species spacer acquisition is common within a natural ecosystem, and that the spacers can inform us of the environmental network of gene exchange. We then demonstrate directly that haloarchaea acquire spacers from the mating partner chromosomes, but also from self-replicons during mating. Finally, we examine the consequences of such spacer acquisition events and show that CRISPR-Cas targeting reduces the frequency of gene exchange via fusion across species, thus restricting horizontal gene transfer across species.

Results

Inter-species targeting is pervasive in halophilic archaea

Previous surveys of CRISPR spacers in bacteria and archaea indicated a dominance for spacers that match viruses known to infect the CRISPR-Cas containing organism or related species [9,15–18]. When we compared the spacers from all halophilic archaea (class Halobacteria) in the CRISPRdb database, to the NCBI database using a sequence similarity search 5.3 % had significant database matches. Surprisingly, most haloarchaeal CRISPR arrays had spacers that matched genes in other haloarchaeal species (Supplementary Table 1). Such cross-targeting spacers were nearly as abundant as spacers matching viral sequences (44% vs. 56% of all spacers with database matches, respectively, Supplementary Tables 2). Haloarchaeal spacers are generally over 30 bases long and hence even near perfect matches are unlikely to occur by chance (see Methods). Most of the spacers with non-viral hits matched genes found on the main chromosomes, while relatively few matched plasmid-encoded genes or transposable elements. Of the chromosomal genes, most were hypothetical genes or had general function prediction only, but several had known housekeeping functions, including genes involved in DNA replication and repair such as those encoding the replicative helicase MCM, and genes encoding the replication initiation protein Orc1/CDC6 (Supplementary Fig. 1 and Supplementary Table 2). While DNA replication and repair genes also occur in selfish elements, the matches in all these cases were to primary chromosomal homologs. Only one strain, *Halorubrum*

lacusprofundi, harbored a spacer that had a significant match to its own genome (matching the *orc1/cdc6* gene). However, this spacer had two mismatches, indicating that it is more likely to have been acquired from a different species rather than being a case of CRISPR autoimmunity, thought to be rare in archaea [19]. In conclusion, spacers that match chromosomal loci in other species are fairly abundant in haloarchaea.

To explore whether this unusual pattern of inter-species targeting is also common in the natural habitat of haloarchaea, we obtained draft genomes of 15 different strains, belonging to four different genera. The strains were isolated from the same small sampling site in Atlit, Israel, less than 100 square meters in area, in the summers of 2012 and 2014. This rocky shore has much evaporation in summer, due to direct sunlight, resulting in small tidal evaporation pools that are hypersaline, and often exhibit a visible salt crust. Since this site is so small and experiences westerly winds daily, cells from one pool can come into contact with those from other pools and potentially even mate by cell fusion [11,20]. Out of these 15 genomes, 11 had both CRISPR arrays and *cas* genes of type I-B CRISPR systems. Of the 1104 spacers in these arrays, only 35 had significant BLASTN matches (3.3%), as is generally the case in both archaea and bacteria [7]. Notably, five of these isolates, belonging to the genera *Haloferax* and *Haloarcula*, had spacers that targeted other strains from the same site, with a total of 16 inter-species spacers (24 when including arrays shared between different strains), 13 of which were perfect matches along the entire spacer length (Table 1, Supplementary Table 3). Most of the spacers matched genes on contigs inferred to be parts of the main chromosomes, rather than plasmids, and some had sequence identity to known house-keeping genes such as ORFs encoding a fatty acid coA ligase and a sugar transporter (Supplementary Tables 4 and 5). Nevertheless, some of these chromosomal targets (4/13) were less than 20Kb away from recombinase genes, indicating that they either target an island or provirus or a chromosomal locus just flanking it (Supplementary Table 6). These results are in agreement with a large survey of spacers in bacteria and archaea, showing that integrated mobile elements represent the most common CRISPR targets [7]. Three *Haloferax* isolates (105R, 109R, 24N), had identical CRISPR arrays despite variable genomic content, and thus were treated as a single CRISPR genotype in subsequent analysis. Ten of the targets were inferred to be part of the main chromosome, while seven matched natural plasmids (Methods). Altogether, in the CRISPR arrays of the

environmental isolates, spacers that matched DNA of other strains outnumbered spacers that matched the genomes of known haloarchaeal viruses in the NCBI database 17 to 10 (Supplementary Table 4), yet it should be considered that archaeal viruses are highly under-represented in the databases.

Table 1. CRISPR spacers of haloarchaeal strains isolated in the summers of 2012/2014 from Atlit. Genus assignment was based on the 16S rRNA and the *polB I* gene sequences.

Strain ID	Year Isolated	Genus of strain	Spacers that match haloarchaea[1]	Spacers that match viruses	Total spacers	*cas* genes presence
19N	2012	*Haloferax*	1	0	**154**	+
48N	2012	*Haloferax*	1*	2	**92**	+
47N	2012	*Haloferax*	3 (1)	1	**70**	+
24N	2012	*Haloferax*	4 (1)	0	**109**	+
105R	2014	*Haloferax*	4 (1)	0	**109**	+
109R	2014	*Haloferax*	4 (1)	0	**109**	+
7R	2014	*Haloarcula*	7 (1)**	0	**111**	+
47R	2014	*Haloarcula*	0	3	**109**	+
120R	2014	*Haloarcula*	0	2	**30**	+
31R	2014	*Halobellus*	0	0	**43**	+
38R	2014	*Halobellus*	0	0	0	-
26R	2014	*Halorubrum*	0	3	**168**	+
8R	2014	*Halorubrum*	0	0	0	-
9R	2014	*Halorubrum*	0	0	0	-
28R	2014	*Halorubrum*	0	0	0	-

[1] In parentheses number of spacers also matching self. * denotes a spacer with a single mismatch to the protospacer (length 37-40bp), ** indicates two such spacers in the respective genome.

Of the 13 cases of perfect spacer-protospacer (target) identity, nearly all were within-genus matches (Fig. 1), while one *Haloferax* spacer matched both a sequence in another *Haloferax* strain and a *Haloarcula* strain. Additionally, a single spacer in *Haloferax* strain 24N only matched a gene in *Haloracula* strain 120R. About a third of the spacers (4/13) that matched chromosomal contigs were in the 3 first (leader-proximal) positions in their respective arrays, and therefore likely to be fairly recent acquisitions, and not necessarily selected for retention [9]. Chromosome-matching

spacers were enriched over 3 fold in these positions compared to their general occurrence in these arrays (hypergeometric p = 0.03), indicating that although they are a small subset of the total spacers in these haloarchaea, they nevertheless constitute a substantial fraction of recent acquisition events.

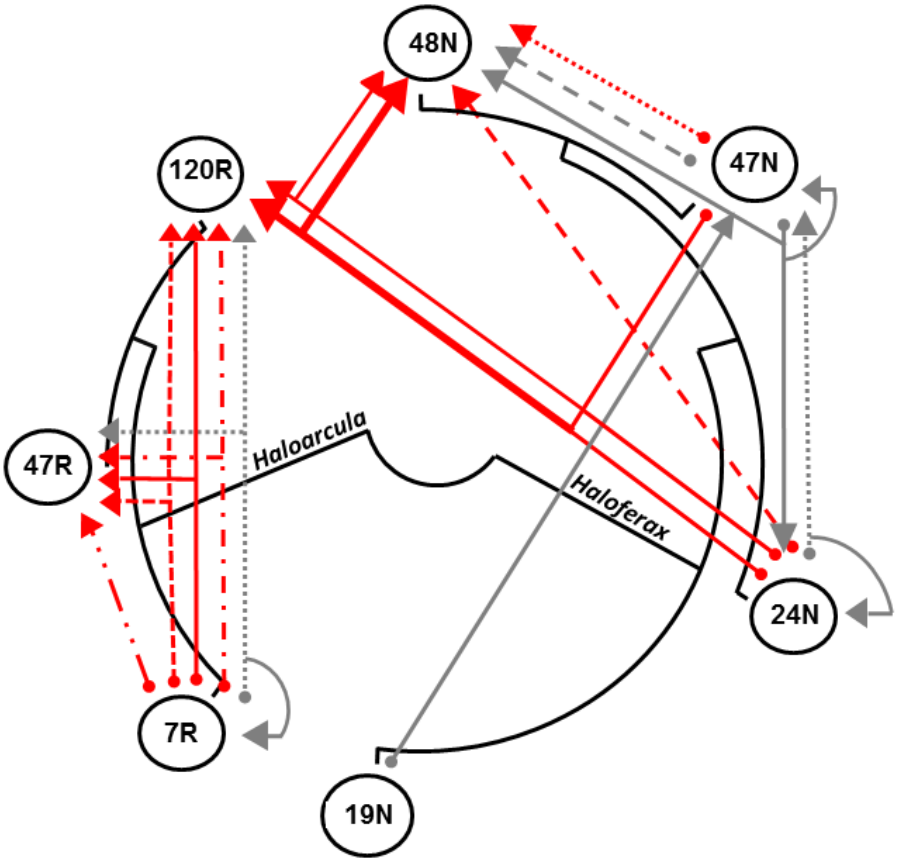


Fig. 1. A Simplified representation of perfect inter-species matches between CRISPR spacers and DNA sequences in the genomes of environmental Atlit isolates. The phylogenetic relationships between isolates, that either target other strains or are targeted by them, are marked in black. Each arrow represents a spacer pointing to the isolate/s it matches. Branching arrows indicate that a spacer has more than one target, including the possibility of self-targeting (circular reference). Gray colored arrows have non-active corresponding interference PAM sequences while red arrows represent spacers with PAMs previously shown to be active in *Haloferax* or *Haloarcula*. Thicker arrows lines indicate a spacer common to both 24N and 47N that targets two genomes. Multiple independent spacers for a given genomic target in the same strain are marked with a different line pattern.

Three of the 13 perfect match spacers were also self-targeting spacers (Fig.1; Supplementary Table 4), i.e. matching a sequence within the same genome that is

outside the CRISPR array. Since having such "auto-immune" spacers is generally considered to be highly deleterious to the organism [19,21–25], it was surprising to observe them in recently isolated strains that are presumably fit. However, target DNA degradation (known as "interference") by type I CRISPR-Cas systems also requires an appropriate protospacer adjacent motif (PAM) sequence. When we inferred the PAM sequences associated with these self-targeting spacers, we observed that all three spacers had PAMs that were previously shown in the same genera as unable to confer efficient interference: in *Haloferax volcanii* [25] the GGC PAM observed in 24N, and the GAT observed in 47N are considered to be inefficient in conferring interference, as was the CCG PAM observed in 7R when previously tested in *Haloarcula hispanica* [26]. In contrast, the most abundant PAM sequence for all the other cross-targeting spacers was TTC, previously shown to be interference-proficient in both *Haloferax* and *Haloarcula* (Supplementary Table 4, [25,26]). Thus, we conclude that these self-targeting spacers are tolerated in the isolates because they are inactive due to the incompatibility of their PAM with the interference complexes of these type I CRISPR-Cas systems (Cascade, [27]). Alignment of the protospacers from isolates and related genomes from the NCBI database showed that these interference-inactive PAMs in those cases are also conserved in genomes that have no such CRISPR self-targeting (Supplementary Fig. 2). Thus, in all likelihood, these inactive PAMs are the product of spacer integration with non-canonical PAMs (see below) rather than subsequent PAM mutations that evade auto-immunity [19].

Acquisition of new inter-species spacers during haloarchaeal mating

The multiple occurrences of inter-species spacers in haloarchaeal genomes raised the question of how such spacers were acquired. One obvious possibility is mating between species by cell fusion [11], during which the entire gene content (both plasmids and chromosomes) of each mating partner is exposed to the other [10], thereby providing an opportunity for acquisition of such cross-species anti-chromosomal spacers. We tested the mating hypothesis experimentally using the two model haloarchaea *H. volcanii* and *Haloferax mediterranei*, which can mate fairly efficiently (only 3.5 times lower than within species mating) [11] despite being quite distant genetically (average nucleotide identity of 86.6% in coding genes, making them

nearly as divergent as *Escherichia coli* and *Salmonella*). These were the first archaeal species that CRISPR arrays were identified in [28] , and both of them possess active CRISPR-Cas systems of subtype I-B, encoded on the large plasmids pHV4 and pHM500 (for *H. volcanii* and *H. mediterranei* respectively) [29–31]. Importantly, spacer acquisition has not been shown for either species, and under normal growth conditions, mRNA levels of cas1 and cas2, the key genes in spacer acquisition are extremely low in both species (13.9 RPKM in *H. volcanii* and 30.7 RPKM in *H. mediterranei*) [32]. After mating *H. volcanii* and *H. mediterranei* and selecting for mated cells we obtained about 200 colonies from which total DNA was extracted. These colonies are not clonal, since they begin from mating products that are heterozygous cells, containing both parental genotypes (chromosomes and plasmids), and later these cells give rise to different recombinant cells that contain chimeric genomes with loci spanning the selectable markers [11]. We then performed PCR on the leader ends of each CRISPR array in the two species (six arrays in *H. mediterranei* and three in *H. volcanii*) followed by gel extraction, secondary PCR, size selection and Illumina amplicon sequencing (240,000-290,000 reads per array per biological repeat, see Methods) to determine which spacers were acquired. We also separately performed shot-gun community sequencing of the same DNA (Supplementary Fig. 3) to gain an estimate of how well were individual CRISPR arrays represented in the mating products. Sequence data analysis revealed substantial acquisition in all arrays in the mating products, with the exception of array E in *H. volcanii* (Supplementary Fig. 4; Supplementary Table 7). Curiously, the vast majority of *H. mediterranei* spacers were derived from its own replicons. Nevertheless, some *H. mediterranei* spacers matched the *H. volcanii* replicons, primarily targeting *H. volcanii*'s plasmid pHV4 (Fig. 2). *H. volcanii* cells acquired more spacers from the *H. mediterranei* chromosome (total spacers=45245, unique spacers=33730) than from all three *H. mediterranei* plasmids combined (total spacers=3114, unique spacers=2176, Fig. 2B), in agreement with findings from the environmental genomes (see above), and unlike previous observations in bacteria, where acquisition was strongly biased toward plasmid DNA [33].

Spacers that target the chromosome are acquired genome-wide

Spacers were acquired by *H. volcanii* from all parts of the *H. mediterranei* chromosome, and included many housekeeping genes, such as those encoding ribosomal proteins, DNA-directed DNA polymerase B2 and DNA polymerase IV, and many others, as seen in the environmental genomes (Fig. 3A; Supplementary Fig. 5A). However, we did observe regions with a higher density of matching spacers (regions that were more than three times higher than neighboring bins) next to putative mobile genetic elements (MGEs, Fig. 3A), which we inferred based on the presence of genes encoding integrases, site-specific recombinases and transposases. One such element, previously referred to as provirus 1, but lacking detectable capsid genes (HHPRED analysis, see Methods), has been previously shown as capable of excising from the genome [31,34]. To test whether this island may excise during our inter-species mating experiments we used inverse PCR to detect its circular (excised) form in the DNA extracted from the same samples that were processed for identification of spacer acquisition. Indeed, we could clearly observe the circular form of that element, which rarely exits the genome of *H. mediterranei* under normal growth conditions [31], but was dominant in the between-species mating experiments (Supplementary Fig. 6). Thus, MGEs that excise from the chromosome may be preferred substrates for spacer acquisition in *Haloferax*. Additional loci of increased acquisition were found close to active CRISPR arrays such as array B and H on the *H. mediterranei* chromosome, and C and D on pHV4 (Fig. 3A; Supplementary Fig. 5), in agreement with previous studies in bacteria [35,36] and archaea [37]. We observed an apparent "no-acquisition zone" in the region between 2899120-2922023 in the *H. mediterranei* genome that hinted that this region has been deleted from the genomes. PCR analysis confirmed the suspicion that this locus had already been deleted in the parental strain WR646 prior to the mating experiments.

In both species, which have roughly similar genome sizes, even when accounting for natural plasmids, many spacers were acquired against self replicons (chromosomes and plasmids). In *H. mediterranei* spacers against self replicons outnumbered those derived from the mating partner about three-fold (Fig. 2), while In *H. volcanii* spacers against self replicons, primarily from pHV4, were approximately as abundant as those

obtained from the *H. mediterranei* replicons, primarily from the major chromosome of the latter species. In terms of acquisition from self replicons both species acquired more spacers from their respective plasmids than from their chromosomes, when normalizing for replicon length (Supplementary Fig. 7).

Interestingly, while *H. volcanii* acquired many spacers from the putative MGEs of *H. mediterranei*, the latter archaeon only showed a hot-spot of acquisition against its own chromosome close to a putative MGE (Fig. 3B; Supplementary Fig. 5B). This implies that the two different CRISPR-Cas systems have different acquisition preferences, even when acquiring from the same replicon.

Within-species mating results in lower levels of spacer acquisition

Given that CRISPR spacer acquisition was induced by inter-species mating, we examined whether mating would also induce acquisition when cells belong to the same species. We therefore performed within-species mating experiments in *H. volcanii*. Experiments comparing between-species (excluding *mediterranei*-derived spacers, from the calculation, Supplementary Table 8) to within-species mating, revealed much reduced spacer acquisition in the *volcanii-volcanii* mating compared to *volcanii-mediterranei* mating. Thus, mating between species leads to subsequent auto-immunity against self-replicons that would not otherwise emerge.

We also tested the effect of "nutritional competence" [38], the ability of *H. volcanii* to take up foreign DNA on spacer acquisition. When *H. volcanii* cells were incubated with high molecular weight *H. mediterranei* DNA, we observed no acquisition of spacers derived from *H. mediterranei*, and a low level of spacers against self-replicons, comparable to that observed in within-species mating (Supplementary Table 8). We thus conclude that nutritional competence is unlikely to be a major source of spacer acquisition in *Haloferax*.

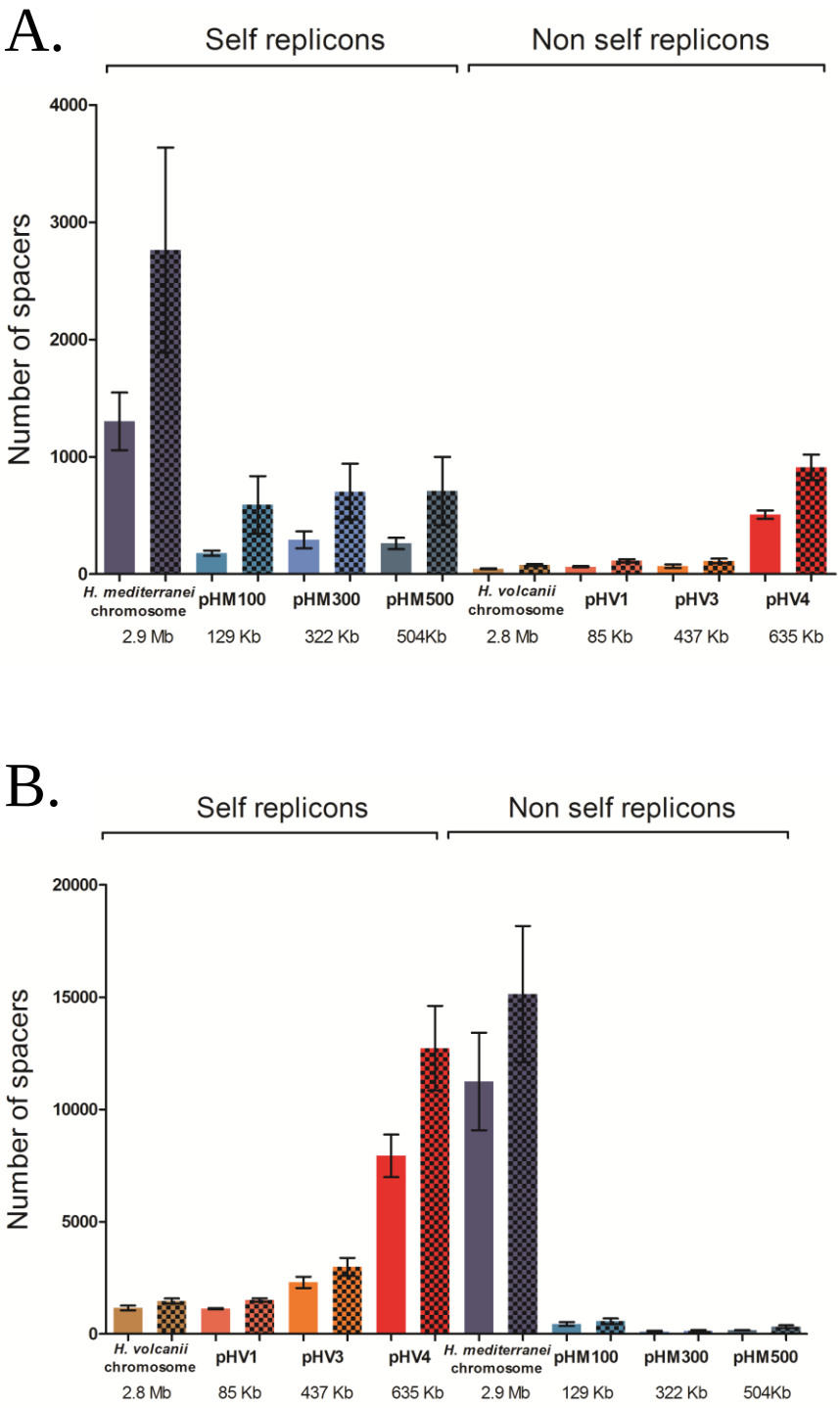
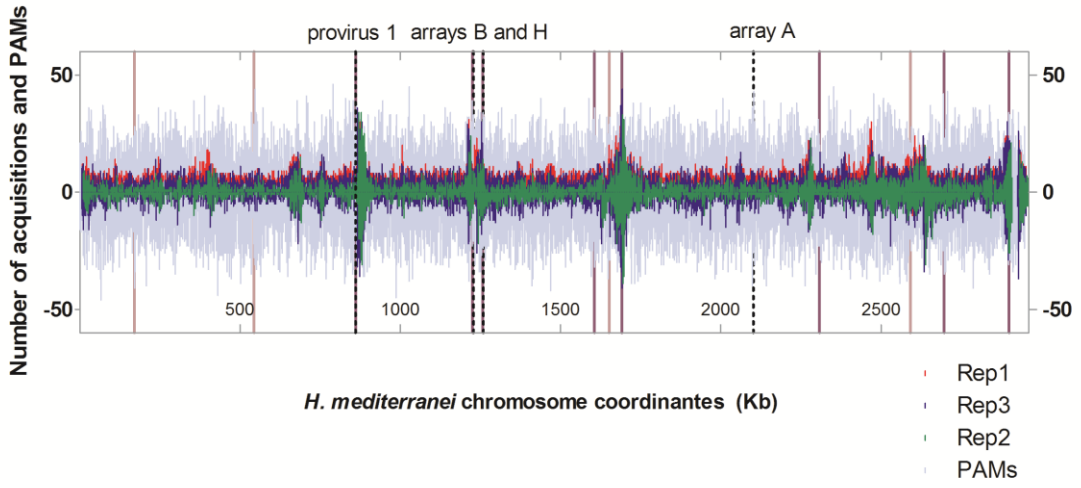


Fig. 2. Number of spacers acquired during mating between *H. volcanii* and *H. mediterranei*. Three independent biological replicates were performed. For each replicon, the mean of unique spacers (no pattern) or total spacers (dotted pattern) is shown. Error bars represent standard error of the mean. **A**. *H. mediterranei* spacer acquisitions. **B.** *H. volcanii* spacer acquisitions.

A.



B.

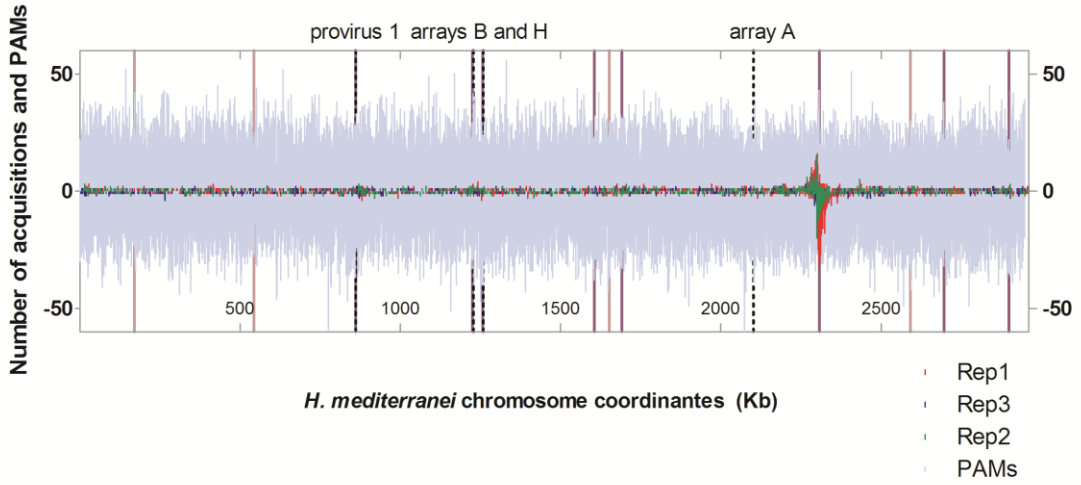


Fig. 3. Spacers acquired from the *H. mediterranei* chromosome by either *H. volcanii* or *H. mediterranei* CRISPR-Cas. The *H. mediterranei* genome was divided into equally sized bins, and the number of unique spacers per bin is represented on the Y axis and colored by biological replicate. PAMs per bin are marked in grey (TAC for *H. volcanii* acquisitions, TTC for *H. mediterranei's*). Dotted lines mark CRISPR arrays and provirus 1. Purple lines mark locations of genes encoding integrases and recombinases and those of genes encoding transposases are denoted by pink lines. **A.** *H. volcanii* PAM signature and spacer acquisition (2458 bp bins). **B.** *H. mediterranei* PAM signature and spacer acquisition (1160 bp bins). Different sized bins were chosen to reflect the fact that the TTC PAM is more than twice as abundant as the PAM in the *H. mediterranei* chromosome.

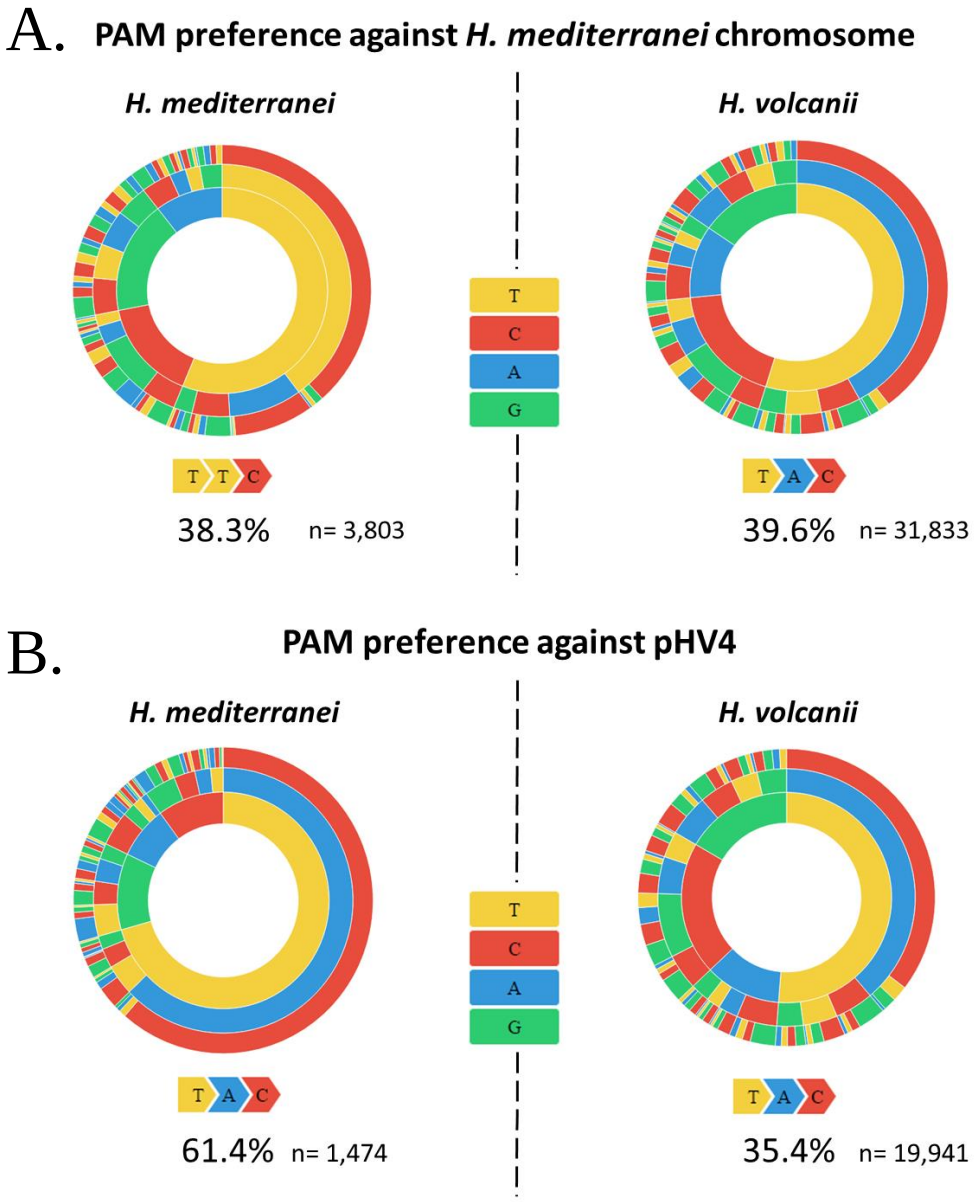


Fig. 4. Protospacer Adjacent Motif (PAM) preference for each of the species against different replicons. Sequences of the three bases upstream for each of the individual unique protospacer were extracted from the acquisition data. The relative abundances of these three-base PAMs were calculated, and are represented in a PAM wheel for each of the species by the spacer target. The favorable PAM (the one with the highest frequency) is marked for each PAM wheel. "n" represents the total number of spacers accounted for in the chart. **A.** PAM wheels for each species based on spacers acquired from *H. mediterranei* chromosome. **B.** PAM wheels for each species based on spacers acquired from *H. volcanii* natural plasmid pHV4.

Protospacer adjacent motifs suggest an interaction between acquisition machineries of the two CRISPR-Cas systems during inter-species mating

The fact that two active CRISPR-Cas systems come into contact during inter-species mating creates an opportunity for them to interact functionally. However, these I-B systems are highly divergent with only 68% and 37% identity between their Cas1 and Cas2 proteins respectively. It is therefore not surprising that their respective leader sequences that are critical for spacer integration also differ (Supplementary Fig. 8). To investigate whether one system could have incorporated spacers produced by the biochemical machinery of the other, we first identified the PAM sequences of each CRISPR-Cas system based on the newly acquired spacer data (Supplementary Fig. 9). The two systems had different PAM signatures: for *H. mediterranei* the preferred PAM sequence was TTC while for *H. volcanii* it was TAC. While the TAC PAM was observed for *H. volcanii* acquisitions from all replicons, the *H. mediterranei* TTC PAM was only observed for spacers acquired from its own replicons (Fig 4; Supplementary Table 9). In contrast, the spacers acquired by *H. mediterranei* from *H. volcanii* replicons instead showed the TAC signature, indicating that they were most likely incorporated into *H. mediterranei* arrays by the ***H. volcanii*** acquisition machinery. In agreement with this conclusion, the pattern of acquisition against pHV4 was also similar between species (Supplementary Fig. 5C and D).

Another interesting feature of these PAMs in both *Haloferax* species was a large fraction of acquired spacers that had non-preferred PAM sequences (Fig 4.), although this fraction was lower when examining only spacers that were observed more than once (non-singletons, Supplementary Fig. 9B). This pattern of acquisition is more noisy than other archaeal CRISPR-Cas systems ([37], [39]), but similar results were obtained in *E. coli* ([33]), and could explain some of the self-targeting spacers with interference-inactive PAMs that we observed in the environmental *Haloferax* isolates (see above).

Acquisition PAMs must be able to also mediate effective interference so that CRISPR-Cas can function as an adaptive immune system. It has been experimentally shown in *H. volcanii* that the CRISPR-Cas system is able to lead to CRISPR-

mediated degradation of artificially transformed plasmids [30]. However, in a screen for nucleotide motifs that could serve as efficient PAMs for DNA degradation in *H. volcanii*, TAC was not identified as an active PAM [30]. Since a specific spacer-PAM combination can sometimes be inactive even when either component is individually active [40], we tested the TAC PAM in *H. volcanii* in an inhibition of transformation assay (Supplementary Table 10). Indeed, a spacer targeting a sequence with the TAC PAM yielded over 100-fold inhibition of plasmid transformation, confirming that this PAM is efficient in mediating degradation of invading DNA by the CRISPR-Cas system of *H. volcanii.*

Inter-species CRISPR targeting has a negative effect on mating success

Our previous experiments clearly showed that spacers can be acquired from another chromosome during inter-species mating by fusion, a mechanism of horizontal gene transfer (HGT) that facilitates the transfer of large plasmids, the emergence of heterozygous fused cells and inter-species recombinant hybrids [11]. However, such spacer acquisition from the other species' genome can reduce the success of future inter-species mating: if a CRISPR-Cas system starts degrading the other genome, this could potentially cause cells to sense DNA damage and separate prematurely, reducing the chances of plasmid exchange and/or recombination between chromosomal loci. To test this hypothesis, we planted a 40 bp sequence that is efficiently targeted by *H. volcanii* CRISPR and mediates interference [25] into the *H. mediterranei* genome generating a targeted *H. mediterranei* strain (Supplementary Fig. 10). We then performed mating assays crossing *H. volcanii* with the targeted *H. mediterranei* strain and as a control we did a parallel experiment crossing *H. volcanii* with an isogenic non-targeted *H. mediterranei strain.* In 9 out of 10 biological replicates we observed a substantial decrease in mating efficiency in comparison to the non-targeted control [nearly 2.5 fold median reduction, $P < 0.002$, Wilcoxon signed paired samples rank test (Fig. 5A)].

To rule out the possibility that mating efficiency is affected by the location of the spacer in the *H. mediterranei* genome, which was in the selectable marker region [*ΔtrpA*(704395)], we created two more targeted strains: one in which the spacer+PAM

sequence was inserted into the *ΔpyrE2*(299911) region, and another that had two specific insertions of that 40 bp sequence, in both *ΔpyrE2* and *ΔtrpA* regions. We observed the same trend of reduction in mating products using both strains in comparison to the non-targeted control (see Fig. 5B). These results clearly indicate that targeting of partner chromosomes by the CRISPR-Cas machinery can reduce HGT by mating across species, even when the selection marker itself is far from the targeted locus.

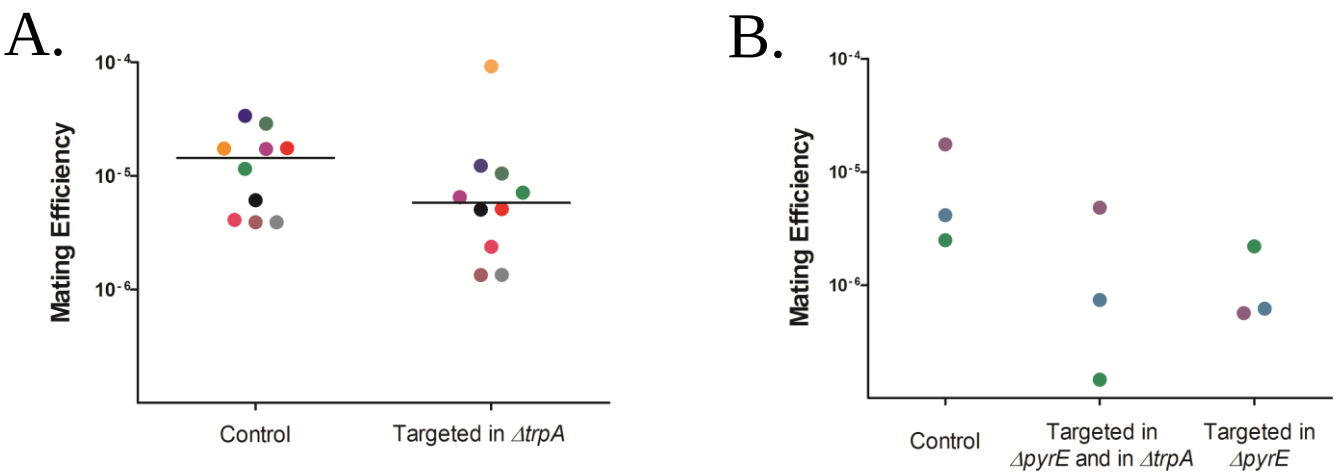


Fig. 5. Inter-species mating is reduced by CRISPR-Cas targeting. *H. volcanii* was mated with either *H. mediterranei* strain WR646 ("Control") or with an isogenic strain engineered to contain a validated *H. volcanii* CRISPR spacer+PAM sequence ("Targeted"), thereby allowing *H. volcanii* to target it during mating (Supplementary Fig. 10). Mating efficiency was calculated as the number of CFUs on the mating plates (mating products) divided by the average number of CFU for each parental strain and are shown for the targeted *H. mediterranei* strain and the targeting-free control. Each independent biological replicate is marked by a different color. **A.** Mating when the target is in the *ΔtrpA* region; the median for each group shown as a line. 9 out of the 10 repeats showed a significant decrease in mating efficiency in comparison to the control. $P < 0.002$, Wilcoxon signed-rank test comparison of paired samples. **B.** Mating with *H. mediterranei* targeted in both *ΔpyrE2* and *ΔtrpA* regions and only in *ΔpyrE2*.

Discussion

The presence of spacers that match chromosomal genes of other archaeal species, in the CRISPR arrays of halophilic archaea represents a record of former genetic interactions with the other species. Here we show experimentally that spacer acquisition from another species' chromosome occurred during inter-species mating, but not during exposure to naked DNA of the other species. Obviously one cannot infer with confidence the type of exposures that environmental strains have undergone. Nevertheless, generalized transduction has never been observed in halophilic archaea, and no genes with sequence similarity those encoding the *Methanococcus voltae* gene transfer agent [41,42] exist in either *H. volcanii* or *H. mediterranei* genome. Our results therefore suggest that the acquisition of spacers from chromosomes of other haloarchaeal species in nature, involves primarily mating.

Nevertheless, the phenomenon of spacers that mostly target chromosomal loci is not unique to archaea, and has also been observed in bacteria, in the genomes of pathogenic *Neisseria* species [43]. Notably, while some of those *Neisseria* spacers matched genomic islands, others were identical to core housekeeping genes, similar to our observations in haloarchaea. *Neisseria* are known to be naturally competent and experience frequent HGT and recombination, in resemblance to *Haloferax* [44], which could provide the opportunity for spacer acquisition from islands as well as other genomic loci, as we have shown here to occur in *Haloferax*. Indeed although spacers in our experiments were acquired from the entire main chromosome, there were obvious hotspots in islands, as observed in *Neisseria,* and more recently in *Pectobacterium atrosepticum* [36]. These results provide direct evidence that spacers can be derived from selfish elements that are incapable of a lytic lifecycle. They also represent strong experimental support for the view that CRISPR-Cas systems play an important role in controlling integrative selfish elements [7], and thereby modulate genome content. This role is complementary to, and may sometimes exceed the importance of, anti-viral defense, especially in the many prokaryotic lineages that have relatively low exposure to viruses but experience frequent DNA exchange.

To date, the regulation of acquisition of spacers has been the least explored aspect of CRISPR activity. Indeed in many bacterial CRISPR-Cas systems, such as I-F systems, it is difficult to disentangle the acquisition from interference, because the key nuclease helicase *cas3* is fused to the acquisition gene *cas2* so that single polypeptide is produced that is involved in both functions [36,45,46]. In contrast, in type I-B systems this is not the case, and consequently there can be robust constitutive interference while the acquisition machinery remains tightly repressed, and such is the case in *Haloferax*. Indeed based on our findings, such regulation is required, since the acquisition machinery in *Haloferax* does not show a preference for non self-replicons, or for plasmids compared to the main chromosome (with the exception of plasmid loci neighboring active arrays in pHV4, Fig. 2), in contrast to what has been observed for bacteria.

A possible explanation for this difference in specificity of acquisition discussed above is that *H. volcanii* like the vast majority of archaea lacks RecBCD, which is responsible for the bias in acquisition that greatly reduces the number of chromosome-derived spacers in *Escherichia coli* [35]. The biases that we did observe seemed to favor acquisition from integrated selfish mobile elements, at least one of which was replicating in a circular form (provirus 1, Supplementary Fig. 6). This bears similarity to recent work in *Pyrococcus furiosus* that has shown acquisition that was biased in favor of rolling circle replication plasmids, but not plasmids that had other replication mechanisms[37]. In both cases spacer acquisition spanned the entire element, without a particular sharp increase at an exposed end, as was reported for viral injection [8].

During inter-species mating *H. volcanii* acquired not only *H. mediterranei*-targeting spacers, but also many self-targeting spacers, primarily from its own plasmids, which can potentially lead to autoimmune targeting of DNA by CRISPR-Cas. CRISPR-mediated deletion events have been shown in engineered *H. volcanii* [47] as well as in bacteria, where CRISPR activity could cause the deletion of genomic islands or parts of them [22,23]. Our results demonstrate that self-targeting by CRISPR can occur naturally in archaea, as a byproduct of inter-species mating. While acquisition from self replicons represents collateral damage incurred due to CRISPR-Cas activity, such accidents may nevertheless have profound effects on genome dynamics. Acquisition

of a spacer targeting an endogenous plasmid gene can lead to DNA degradation and result in either loss of the targeting activity by mutational events, or in the deletion of the target region [19,30]. Which of these scenarios dominates will probably be determined by the cost or benefit that this plasmid provides in a given environment [48]. Such semi-random deletion processes will yield a population of cells that carry plasmids differing in their gene content, thereby increasing the genotypic variation within the meta-population. Furthermore, since the costs and benefits of plasmid-encoded genes will vary with environmental fluctuations, this process will increase the chances that a transiently fit genotype will emerge, and thus will benefit the overall population fitness.

Acquisition from the mating partner's chromosome, could affect horizontal gene transfer (HGT) between species. We show that once a species' CRISPR-Cas system effectively targets another's genome, the frequency of productive mating events between them drops substantially. While the 2.5 fold reduction we observed may seem small compared to previously reported effects of CRISPR-Cas on HGT[49], it is noteworthy that euryarchaeal DNA is packaged with histones and less accessible to CRISPR-Cas than viral or naked DNA. Indeed histones have been recently been shown to protect the DNA of another euryarchaeon, *Methanocaldococcus jannaschii*, from degradation by the arogonaut defense system [50]. Nevertheless, the decrease in mating that we observed is similar to the reduction in mating efficiency noted when cross-species mating (i.e. mating between *H. volcanii* with *H. mediterranei* wild type cells) is compared to a situation when the mating partners are both *H. volcanii* (3.5 fold median reduction)[11]. Importantly, The effect of CRISPR-Cas that we observed reduced gene exchange *globally* and not just *locally* at the CRISPR-targeted site: both our selectable marker loci were not close to the targeting site. One may speculate that during mating that involves CRISPR targeting of the chromosome, the targeted cell may simply sense damage to DNA and detach prematurely, resulting in fewer successful HGT events.

In summary, spacer acquisition from chromosomes, whose targets may occasionally be genuine housekeeping genes, can be a common side effect of CRISPR-Cas activity that is primarily directed against selfish elements, especially islands. Such accidental

acquisition however, could then have global impacts on gene exchange, and increase genetic separation between lineages, contributing to speciation.

Methods

Identification of haloarchaeal CRISPR spacers that match sequences in sequencedatabases

All 1161 high confidence (confirmed) spacers from pre-existing haloarchaeal genomes were downloaded from the CRISPRdb website (http://crispr.u-psud.fr/crispr/database, see [51] and compared using NCBI BLASTN (last GenBank version-18/6/16) to Halobacteria (taxid:183963) and viruses (taxid:10239) in the NCBI database. Self-hits (hits of spacers against themselves within the CRISPR array context) were filtered out by removing all 100% identity hits where the organism as well as the locus of the spacer and its matching sequence were the same. We examined only hits that met all the following criteria: Coverage ≥ 0.5, Score ≥ 40 bits and E-value ≤ 0.001.

Analysis of spacers in environmental isolates from the Atlit seashore

Isolates were collected from evaporation puddles, less than 100 square meters in area, on the coast of Atlit, Israel, in the summers of 2012 and 2014. This rocky shore has much evaporation in summer, due to direct sunlight, resulting in small tidal evaporation pools that are hypersaline, and often exhibit a visible salt crust. Since this site is so small and experiences westerly winds daily, cells from one pool can come into contact with those from other pools and potentially even mate by cell fusion [11,20]. Isolates were sequenced by Illumina 2×250 base paired end whole genome sequencing. Genus identity was determined according to 16S rRNA and the *polB* genes sequences. Raw reads were first trimmed with Cutadapt v1.9.1[52] to remove adaptor sequences and bases with a Phred score lower than 20. Genomes were assembled using SPAdes v3.7.0 [53] with kmer sizes 21,33,55,77,99, and 127. Assembly quality was checked using Quast v2.3[54], coding sequences and annotations were predicted using Prokka v1.11 [55] ignoring contigs that were shorter than 200 base pairs. CRISPR arrays were identified using the CRISPR Recognition Tool (CRT) v1.1. [56] and CRISPR-finder from the CRISPRdb website [51]. We extracted all confirmed array locations and spacers for each isolate. Cross-targeting spacers were identified using a BLASTN search against the isolates combined genome

files, followed by manual elimination in cases of shared spacers or CRISPR arrays using the CRISPR array coordinates in the targeted isolate. We also searched traces of degenerate repeat sequences (up to 10 mismatches) in proximity to the protospacer in order to eliminate cases of "false hits". We designated "self" targeting spacers, those spacers from confirmed arrays that had match to their own genome in locations that did not map to other confirmed/hypothetical CRISPR arrays. Protospacer location was estimated as chromosomal in cases where conserved housekeeping genes were present on the same contig (DNA and RNA polymerase subunits and genes encoding rRNA and core ribosomal proteins), otherwise the location was assumed to be plasmid. Isolates' confirmed spacers were also compared using NCBI BLASTN to all viruses in the NCBI database with the criteria: Coverage ≥ 0.5, Bit Score ≥40 and E-value ≤ 0.001. All "self"-targeted isolates genomes had fully intact open reading frames of all *cas* genes known to be involved in degrading targeted DNA ("interference") (i.e. *cas*3, 5, 6, 7 and 8). Three *Haloferax* isolates (105R, 109R, 24N), had identical CRISPR arrays despite variable genomic content, and thus were treated as a single CRISPR genotype in subsequent analysis.

Culture Conditions

H. volcanii and *H. mediterranei* cells were routinely grown as described in [57].

***Haloferax* strain construction**

Strain construction were performed according to the protocol described in [58,59].

Mating experiments

The mating experiments were conducted using an *H. volcanii* strain lacking the ability to synthesize thymidine H729 (Δ*hdrB [*2754021]) and an *H. mediterranei* strain that is unable to synthesize tryptophan and uracil WR646 (Δ*trpA* [704395]*,* Δ*pyrE2* [299911], Supplementary Table S11**)**. Liquid cultures of both parental strains were grown to an O.D600 of ~1.8. The parental strains were then mixed in 1:1 ratio and applied to a nitrocellulose 0.45µm filters using a Swinnex 25mm filter holder. The filter with the mating products was transferred to a rich medium plate (Hv-YPC with thymidine) for 24 hours for phenotypic expression. The cells were then re-suspended and washed in Hv-Ca (*Haloferax volcanii* casamino acids) broth before plated on Hv-Ca media containing tryptophan (*H. volcanii* and *H. mediterranei* mating products were selected using the *pyrE2* and *hdrB*

chromosomal markers).The mating "within species" in *H. volcanii* was performed using H729 (Δ*hdrB* [2754021]) and WR536(Δ*trpA* [302281], Δ*pyrE2* [301751]) selecting on the *pyrE2* and *hdrB* markers.

Nutritional competence experiments

H. volcanii strain H729 cells were grown for a week in YPC-HV broth containing extracted DNA from *H. mediterranei* strain WR646 (50ng/µl). DNA from that culture was then extracted and used for detection of new spacer's in arrays C and D of *H. volcanii* as described below.

Community DNA sequencing of mating products

About 200 colonies of mating products (derived from *H. volcanii* strain H729 and *H. mediterranei* strain WR646) from each of three independent mating experiments were suspended in Hv-Ca broth media prior to DNA extraction using the DNA spooling protocol described in [58]. Purified DNA was sent for metagenomic sequencing using the Nextera XT protocol and the Illumina NextSeq500 sequencing platform at the Center for Genomic Research, University of Illinois, Chicago, USA. After trimming of adaptors and low quality bases by Trimomatic [60], between 1.8 to 2.1 million high-quality sequence reads were obtained for each biological replicate and matched, using blastn, to reference sequences of *H. volcanii* and *H. mediterranei*, including all their natural plasmids; to improve accuracy, the Blast e-value threshold used for mapping reads was set to $1e^{-60}$. Reads mapping to more than one locus were ignored. Reads counts were normalized per kb of sequence.

Detection of acquisition of new spacers

About 200 colonies of mating products (using *H. volcanii* strain H729 and *H. mediterranei* strain WR646 as the parental strains for *H. volcanii-H. mediterranei* mating, WR510 and UG453 for *H. mediterranei* within-species mating, and H729 and WR536 for *H. volcanii* within-species mating) were suspended together in Hv-Ca broth media before extracting their DNA using the DNA spooling protocol as described previously [58]. The extracted DNA was used as template for PCR for both species CRISPR arrays using specific primers amplifying the region between the leader and the third spacer in the array for arrays A-H,

and the region between leader and the end of the first spacer in array I (see primers list Supplementary Table 12). When analyzed by agarose gel electrophoresis, longer PCR products indicate new spacer-repeat acquisitions. To obtain visible acquisition bands in the agarose gel, we extracted the approximated elongated length region from the gel, isolated DNA and amplified the fragment through another cycle of PCR. New acquisition events could then be detected via the presence of a higher band. PCR products were then sent for processing and Illumina amplicon sequencing (240,000-290,000 reads per array per biological repeat) at the Center for Genomic Research, University of Illinois, USA. Briefly, the elongated PCR product was enriched using Ampure beads size-selection; sample specific barcodes and Illumina adaptors were added by PCR; and the resulting products were purified, pooled, and paired-ends sequenced on a MiSeq Illumina platform. Notably, even after these consecutive steps of size selection many reads still represented amplicons derived from no acquisition amplifications.

Initial data processing

Paired-end raw Illumina reads were quality-filtered (Q>20) and merged using PEAR (paired-end read merger, [9]), yielding, for most samples, 240000-290000 high-quality sequences; samples which yielded over 300000 seqs per sample were subsampled randomly, using VSEARCH, to 280000 seqs/sample. Biological replicates within each array were then converted to a single fasta file using QIIME's multiple_split_libraries_fastq.py script [61], followed by de-replication, abundance sorting and clustering (99% threshold) using VSEARCH [62]. Pairwise identity at the clustering step was defined as [matching columns]/[alignment length] (set in VSERACH as –iddef 1), which ensures terminal gaps are not ignored and sequences of different length do not cluster together. The length distribution of the clustered sequences is similar to that of the raw sequences; arrays A-H show a peak at 250 bp (size of the original fragment, containing 3 repeats), with progressively smaller peaks at 320 bp (corresponding to one new acquisition) and 390 bp (2 new acquistions). For array I, the original fragment size was 320 bp (since the forward primer, for technical reasons, was located at the beginning of the leader sequence), with corresponding peaks at 390 and 460 bp. The raw reads were then mapped back to the clusters, again with a 99% identity threshold, to create a table presenting the abundance of each cluster in each biological repeat.

Spacer extraction

The centroid sequence of each cluster was used for extraction of acquired new spacer sequences. To identify true acquisition events while excluding DNA rearrangement events (which may also result in an extended PCR product), a custom made R script based on the 'Biostrings' and 'tools' R packages was used to count the number of repeats in each centroid sequence, allowing up to 2 mismatches per repeat . Fragments containing more than 3 repeats (for arrays A-H) or more than 1 repeat (array I) were tagged as putative acquisitions, and the sequence between the 2 repeats closest to the leader end was extracted. "False positives", which are the result of rearrangement of spacers within or between arrays rather than canonical acquisition, were eliminated by screening the extracted putative spacers against all original CRISPR arrays from both species, allowing up to 5 mismatches per spacer; spacers with matches in existing CRISPR arrays were excluded from further analysis.

To facilitate comparison of CRISPR activity between arrays, we also counted the number of sequences containing the original number of repeats (3 for arrays A-H; 1 for array I). All spacer sequences, positions and PAM's are provided (Supplementary Tables 13 and 14).

Mapping new spacers to genomic location

In order to establish the protospacer location for each new acquisition, we used the blastn-short program [63] at an E-value of .0001 against a file containing genome sequences of both *H. mediterranei* and *H. volcanii*, including their natural plasmids. Blast results were refined using Custom made R scripts based on the 'stringr' package. In brief, no more than 3 mismatches between the spacer and the protospacer were allowed; in cases of multiple matches, the location with the highest score was preserved (while conserving abundance information); and spacers with equally high scores across multiple locations were excluded. Furthermore, spacers that were aligned only partially were removed when the alignment length was shorter than the total spacer length by five or more bases.

Quantification of spacer acquisition events

Using the mapping approach described above we quantified total acquisition events for each spacer and the mean number of events across biological replicates for each spacer was calculated. 4 spacers that showed a highly aberrant pattern when comparing the 3 biological replicates (abundance in one sample more than 1000-fold higher than in the other 2 samples) were corrected for by replacing the aberrantly high value with the average counts of that spacer in the other 2 biological replicates.

PAM determination

The ten upstream bases from each unique protospacer were extracted, while adjusting for the rare cases where there was a misalignment within the first five base pairs of the spacer. We observed that only the final three bases contained a non-random PAM sequence. The relative abundance of these three-base PAMs was calculated, and represented in a PAM-wheel [64] using SunburstR, Yaml, and Rcpp R libraries. In cases where a single unique spacer was aligned in multiple locations, (usually corresponding to transposable element repeat sequences that appear multiple times throughout the genome), those matches were removed from the quantitative analysis, since in those cases the protospacer they were acquired from cannot be unambiguously determined.

Annotation of newly acquired spacer targets

The location of protospacers was annotated for each of the acquired spacers using the publically available *H. volcanii* [65] and *H. mediterranei* [66] genomes [NCBI GenBank accession files [67]] CP001868-1871 and CP001953-1957.

Bioinformatic analysis of provirus 1

Proteins sequences from the predicted island regions (HFX_0898-0929) were submitted to the HHpred server for remote homology detection and structure prediction [68] to rule out proteins with significant similarity to known capsid genes. Such approaches have been shown to be sensitive in detecting novel viral capsid proteins in archaeal genomes [69].

Plasmid invader tests

The *H. volcanii* strain H119 was transformed with the invader plasmid pTA409-PAM28-P1.1, which carries the PAM TAC upstream of spacer 1 of CRISPR locus P1 [30]. As a control *H. volcanii* cells were transformed with the vector pTA409. Plasmids were passaged through *E. coli* GM121 cells to avoid methylation and subsequently introduced into *H. volcanii* using the PEG method. The transformations were repeated four times and plated on Hv-Ca plates without uracil to ensure selection for the plasmids. Transformations with at least a 100-fold reduction in transformation rate are defined as successful interference reactions [30].

Data availability

Genomes for Haloarchaeal environmental isolates are available at Genbank, under the following accession numbers: PSYS00000000, PSYT00000000, PSYU00000000, PSYV00000000, PSYW00000000, PSYX00000000, PSYY00000000, QEQI00000000, QEQJ00000000, QPLN00000000, QPLO00000000, QPLP00000000, QPLQ00000000, QPLR00000000, QPLS00000000, QPLT00000000, QPLU00000000, QXIJ00000000, QXIK00000000, and QPLV00000000. All scripts relating to new spacer acquisition and analysis had been deposited in https://github.com/leahfa/CrispR-analysis.

Author Contributions

U.G. and I.T.-G conceived the study; U.G. I.T-G, and A.M. designed experiments; S.M.S. assembled and annotated genome sequences; I.T.-G, A.N., N.A-P designed and constructed strains; I.T.-G, S.Y., K.E., Y.S., A-E.S., and M.Z. performed experiments; L.R, S.M, I.T.G, and U.G. analyzed data; U.G. and I.T.-G. wrote the manuscript, L.R., S.M.S. and A.M. commented and made critical revisions to the manuscript.

Competing interests

The authors declare no competing interests.

Correspondence should be addressed to Uri Gophna urigo@tauex.tau.ac.il

Acknowledgments

The authors wish to thank Rotem Sorek and Anat Herskovits for their helpful comments and insights, and Hua Xiang for providing sequence data and provirus annotations. The authors thank Stefan Green of the University of Illinois at Chicago, for his continued expert help in challenging sequencing projects and Eugene Koonin (NIH) for helpful discussions. Funding: Deutsche Forschungsgemeinschaft [MA1538/16-2]; Israel Science Foundation [535/15]; Binational science Foundation [2013061] and partial support by the Constantiner Institute; European Research Council (grant ERC-AdG 787514).

References and Notes:

1. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–12 (2007).

2. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1390–1400 (2008).

3. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322,** 1843–5 (2008).

4. Shah, S. A., Hansen, N. R. & Garrett, R. A. Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochem. Soc. Trans.* **37,** 23–8 (2009).

5. Held, N. L., Herrera, A., Quiroz, H. C. & Whitaker, R. J. CRISPR associated diversity within a population of Sulfolobus islandicus. *PLoS One* **5,** (2010).

6. Ratner, H. K., Sampson, T. R. & Weiss, D. S. I can see CRISPR now, even when phage are gone: a view on alternative CRISPR-Cas functions from the prokaryotic envelope. *Curr. Opin. Infect. Dis.* **28,** 267–74 (2015).

7. Shmakov, S. A. *et al.* The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *MBio* **8,** e01397-17 (2017).

8. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544,** 101–104 (2017).

9. Brodt, A., Lurie-Weinberger, M. N. & Gophna, U. CRISPR loci reveal networks of gene exchange in archaea. *Biol. Direct* **6,** 65 (2011).

10. Rosenshine, I., Tchelet, R. & Mevarech, M. The mechanism of DNA transfer in the mating system of an archaebacterium. *Science* **245,** 1387–1389 (1989).

11. Naor, A., Lapierre, P., Mevarech, M., Papke, R. T. & Gophna, U. Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr. Biol.* **22,** 1444–1448 (2012).

12. Schleper, C., Holz, I., Janekovic, D., Murphy, J. & Zillig, W. A multicopy plasmid of the extremely thermophilic archaeon Sulfolobus effects its transfer to recipients by mating. *J. Bacteriol.* **177,** 4417–26 (1995).

13. Kuwabara, T. *et al.* Thermococcus coalescens sp. nov., a cell-fusing hyperthermophilic archaeon from Suiyo Seamount. *Int. J. Syst. Evol. Microbiol.* **55,** 2507–2514 (2005).

14. Burstein, D. *et al.* New CRISPR–Cas systems from uncultivated microbes. *Nature* **542,** 237–241 (2016).

15. Haft, D. H., Selengut, J., Mongodin, E. F., Nelson, K. E. & White, O. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Comput. Biol.* **1,** e60 (2005).

16. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151,** 653–663 (2005).

17. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).

18. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60,** 174–182 (2005).

19. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26,** 335–340 (2010).

20. Chimileski, S., Franklin, M. J. & Papke, R. Biofilms formed by the archaeon Haloferax volcanii exhibit cellular differentiation and social motility, and facilitate horizontal gene transfer. *BMC Biol.* **12,** 65 (2014).

21. Yosef, I., Goren, M. G., Kiro, R., Edgar, R. & Qimron, U. High-temperature protein G is essential for activity of the Escherichia coli clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 20136–41 (2011).

22. Vercoe, R. B. *et al.* Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. *PLoS Genet.* **9,** e1003454 (2013).

23. Selle, K., Klaenhammer, T. R. & Barrangou, R. CRISPR-based screening of genomic island excision events in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 8076–81 (2015).

24. Li, Y. *et al.* Harnessing Type I and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Res.* **44,** e34–e34 (2016).

25. Fischer, S. *et al.* An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.* **287,** 33351–33365 (2012).

26. Li, M., Wang, R. & Xiang, H. Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Res.* **42,** 7226–35 (2014).

27. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).

28. Mojica, F. J. M., Juez, G. & Rodriguez-Valera, F. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol. Microbiol.* **9,** 613–621 (1993).

29. Brendel, J. *et al.* A complex of cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (CRISPR)-derived RNAs (crRNAs) in haloferax volcanii. *J. Biol. Chem.* **289,** 7164–7177 (2014).

30. Fischer, S. *et al.* An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.* **287,** 33351–63 (2012).

31. Li, M. *et al.* Characterization of CRISPR RNA biogenesis and Cas6 cleavage-mediated inhibition of a provirus in the haloarchaeon Haloferax mediterranei. *J. Bacteriol.* **195,** 867–875 (2013).

32. Artieri, C. G. *et al.* Cis-regulatory evolution in prokaryotes revealed by interspecific archaeal hybrids. *Sci. Rep.* **7,** 3986 (2017).

33. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40,** 5569–76 (2012).

34. Yang, H. *et al.* Activation of a dormant replication origin is essential for Haloferax mediterranei lacking the primary origins. *Nat. Commun.* **6,** 8321 (2015).

35. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520,** 505–510 (2015).

36. Staals, R. H. J. *et al.* Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. *Nat. Commun.* **7,** 12853 (2016).

37. Shiimori, M. *et al.* Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx839

38. Chimileski, S., Dolas, K., Naor, A., Gophna, U. & Papke, R. T. Extracellular DNA metabolism in Haloferax volcanii. *Front. Microbiol.* **5,** 57 (2014).

39. Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of *S ulfolobus*. *Mol. Microbiol.* **91,** 900–917 (2014).

40. Maier, L.-K. *et al.* Essential requirements for the detection and degradation of invaders by the Haloferax volcanii CRISPR/Cas system I-B. *RNA Biol.* **10,** 865–74 (2013).

41. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* **10,** 472 (2012).

42. Bertani, G. Transduction-like gene transfer in the methanogen Methanococcus voltae. *J. Bacteriol.* **181,** 2992–3002 (1999).

43. Zhang, Y. *et al.* Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. *Mol. Cell* **50,** 488–503 (2013).

44. Papke, R. T., Koenig, J. E., Rodríguez-Valera, F. & Doolittle, W. F. Frequent recombination in a saltern population of Halorubrum. *Science (80-.).* **306,** 1928–1929 (2004).

45. Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E5122–E5128 (2017).

46. Patterson, A. G., Chang, J. T., Taylor, C. & Fineran, P. C. Regulation of the Type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. *Nucleic Acids Res.* **43,** 6038–6048 (2015).

47. Stachler, A.-E. *et al.* High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. *Nucleic Acids Res.* **45,** 5208–5216 (2017).

48. Hullahalli, K., Rodrigues, M. & Palmer, K. L. Exploiting CRISPR-Cas to manipulate Enterococcus faecalis populations. *Elife* **6,** (2017).

49. Zhang, Y. *et al.* Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. *Mol. Cell* **50,** 488–503 (2013).

50. Zander, A. *et al.* Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. *Nat. Microbiol.* **2,** 17034 (2017).

51. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8,** 172 (2007).

52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

53. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

54. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).

55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

56. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).

57. Allers, T., Barak, S., Liddell, S., Wardell, K. & Mevarech, M. Improved strains and plasmid vectors for conditional overexpression of His-tagged proteins in Haloferax volcanii. *Appl. Environ. Microbiol.* **76,** 1759–1769 (2010).

58. Allers, T., Ngo, H. P., Mevarech, M. & Lloyd, R. G. Development of Additional Selectable Markers for the Halophilic Archaeon Haloferax volcanii Based on the leuB and trpA Genes. *Appl. Environ. Microbiol.* **70,** 943–953 (2004).

59. Bitan-Banin, G., Ortenberg, R. & Mevarech, M. Development of a Gene Knockout System for the Halophilic Archaeon Haloferax volcanii by Use of the pyrE Gene. *J. Bacteriol.* **185,** 772–778 (2003).

60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

61. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).

62. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4,** e2584 (2016).

63. Tom Hall. BioEdit Sequence Alignment Editor for Windows 95/98/NT/XP/Vista/7. (2013). Available at: http://www.mbio.ncsu.edu/BioEdit/bioedit.html. (Accessed: 9th February 2017)

64. Leenay, R. T. *et al.* Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol. Cell* **62,** 137–147 (2016).

65. Hartman, A. L. *et al.* The Complete Genome Sequence of Haloferax volcanii DS2, a Model Archaeon. *PLoS One* **5,** e9605 (2010).

66. Han, J. *et al.* Complete genome sequence of the metabolically versatile halophilic archaeon Haloferax mediterranei, a poly(3-hydroxybutyrate-co-3-hydroxyvalerate) producer. *J. Bacteriol.* **194,** 4463–4 (2012).

67. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).

68. Zimmermann, L. *et al.* 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

69. Makarova, K. S. *et al.* Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles* **18,** 877–93 (2014).

70. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497–3500 (2003).

71. Abu-Qarn, M., Eichler, J. & Sharon, N. Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea. *Curr. Opin. Struct. Biol.* **18,** 544–550 (2008).

72. Delmas, S., Shunburne, L., Ngo, H.-P., Allers, T. & Myllykallio, H. Mre11-Rad50 Promotes Rapid Repair of DNA Damage in the Polyploid Archaeon Haloferax volcanii by Restraining Homologous Recombination. *PLoS Genet.* **5,** e1000552 (2009).

Author Contributions

U.G. and I.T.-G conceived the study; U.G. I.T-G, and A.M. designed experiments; S.M.S. assembled and annotated genome sequences; I.T.-G, A.N., N.A-P designed and constructed strains; I.T.-G, S.Y., K.E., Y.S., A-E.S., and M.Z. performed experiments; L.R, S.M, I.T.G, and U.G. analyzed data; U.G. and I.T.-G. wrote the manuscript, L.R., S.M.S. and A.M. commented and made critical revisions to the manuscript.

Acknowledgments

The authors wish to thank Rotem Sorek and Anat Herskovits for their helpful comments and insights, and Hua Xiang for providing sequence data and provirus annotations. The authors thank Stefan Green of the University of Illinois at Chicago, for his continued expert help in challenging sequencing projects and Eugene Koonin (NIH) for helpful discussions. Funding: Deutsche Forschungsgemeinschaft [MA1538/16-2]; Israel Science Foundation [535/15]; Binational science Foundation [2013061] and partial support by the Constantiner Institute; European Research Council (grant ERC-AdG 787514).

Supplementary Information for

Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation

Authors: Israela Turgeman-Grott[1], Shirley Joseph[1], Sam Marton[1], Kim Eizenshtein[1], Adit Naor[1†], Shannon Soucy[2], Aris-Edda Stachler[3], Yarden Shalev[1], Mor Zarkor[1], Leah Reshef[1], Neta Altman-Price [1], Anita Marchfelder[3], and Uri Gophna[1*].

Affiliations:

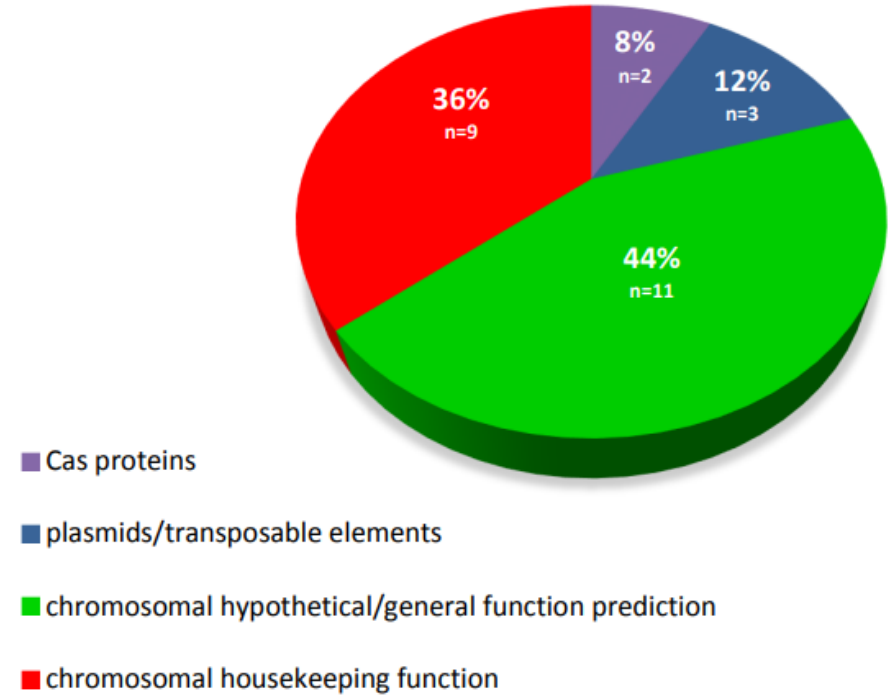
[1] Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978-01, Israel.

[2] Department of Biological Sciences, Dartmouth College, 78 College Street, Hanover, New Hampshire, 03766, USA.

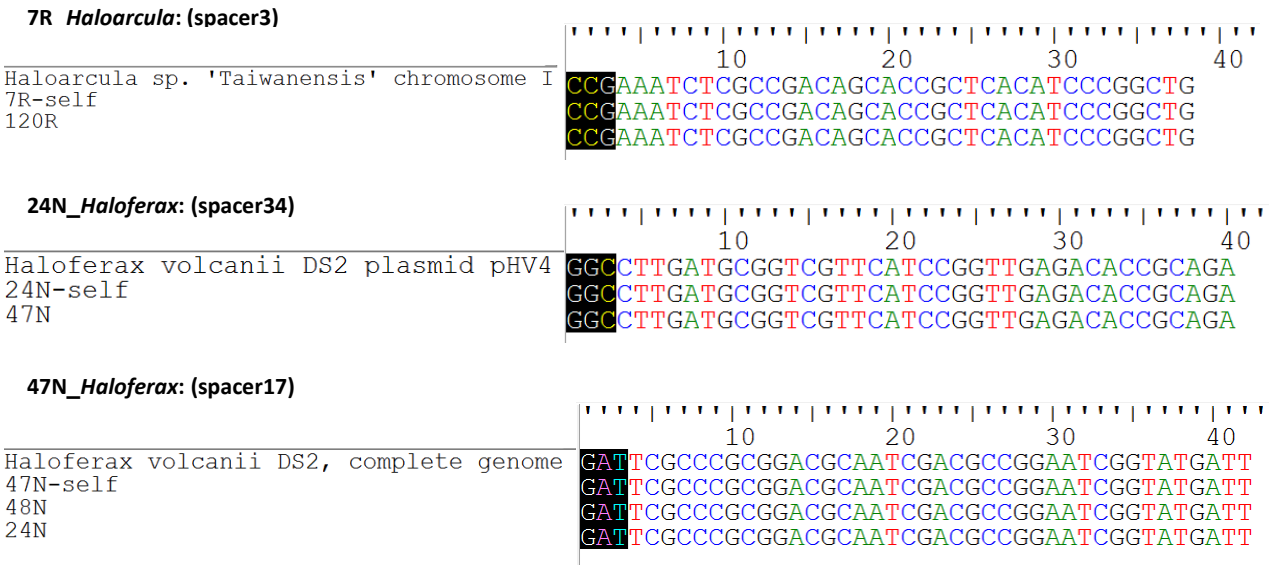
[3] Department of Biology II, Ulm University, 89069 Ulm, Germany.

[†]Present address: Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305.

Supplementary Figures

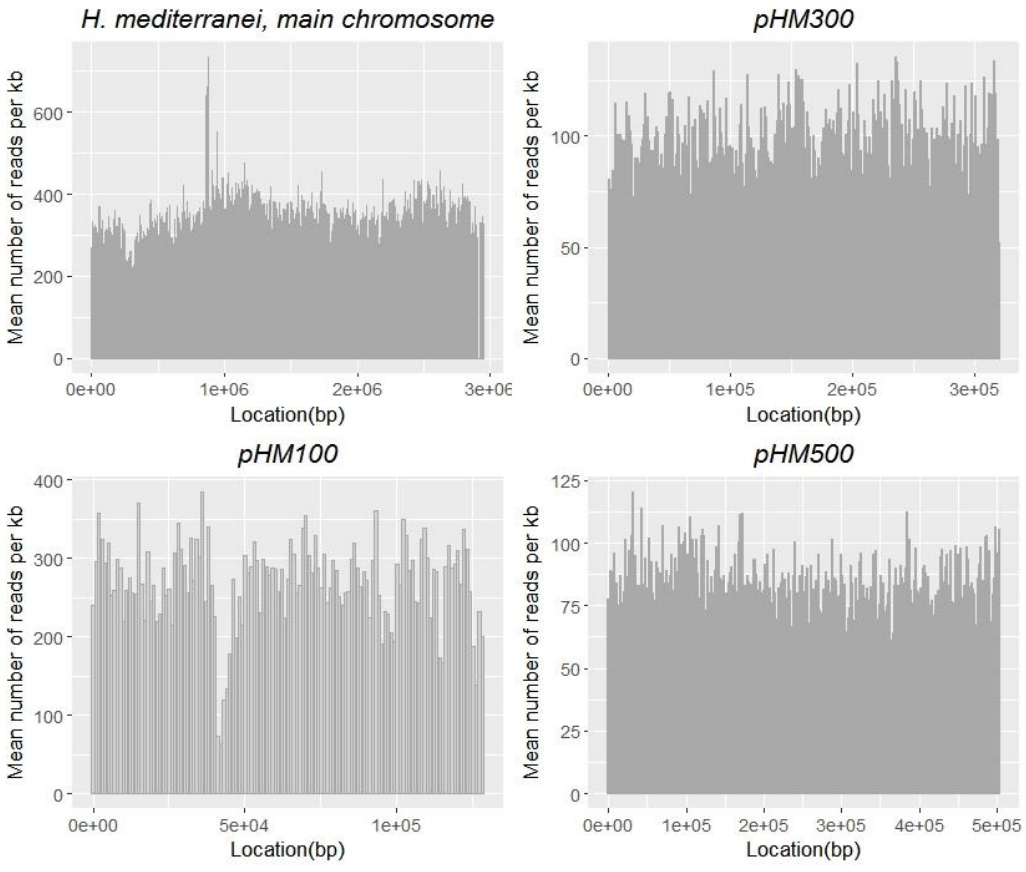


Supplementary Figure 1: Breakdown of haloarchaeal CRISPR spacers matching targets within other haloarchaeal species. 26 confirmed spacers from CRISPRdb [1] that had significant matches to archaeal genomes are shown (see Materials and Methods). A double appearance of the same spacer in a genome is regarded as a single occurrence of that spacer in this analysis.

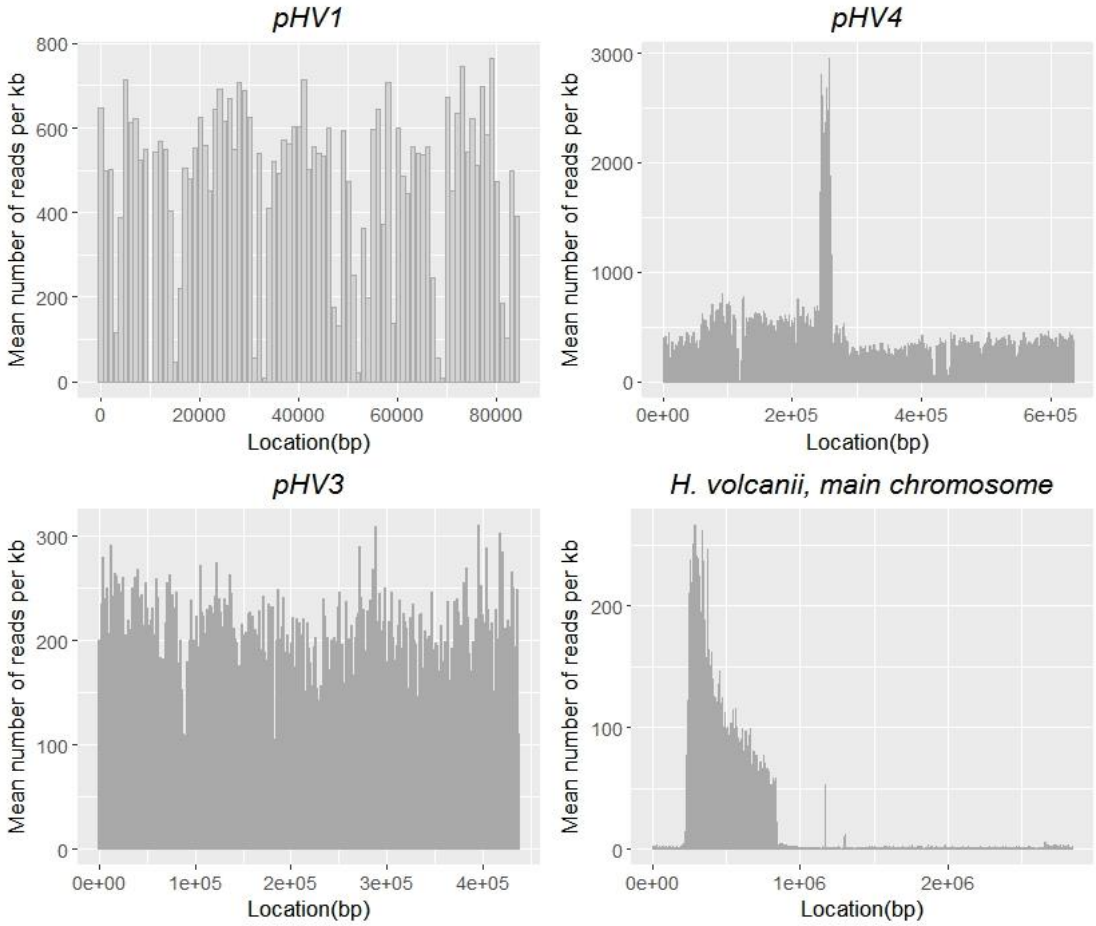


Supplementary Figure 2: Alignment of the protospacer sequences of the self-targeting spacers from Atlit isolates. Three self-targeting spacers were aligned with their respective perfect match protospacers: self chromosomal target, other isolates and database match. The PAM sequence is shaded.

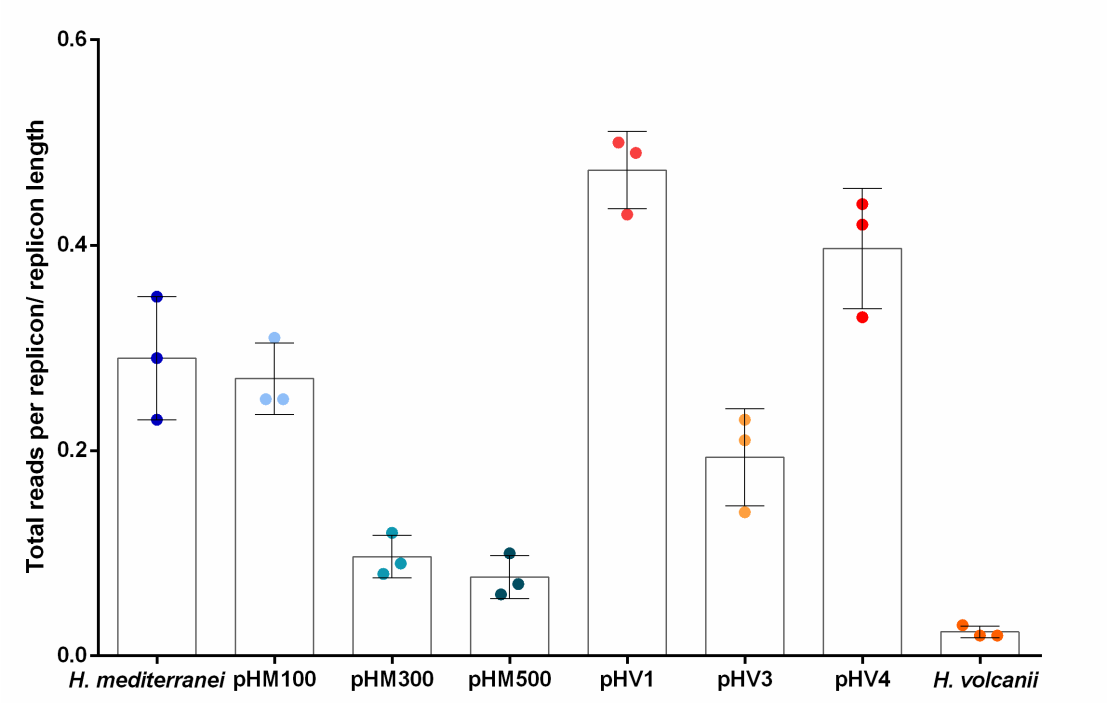
a.



b.

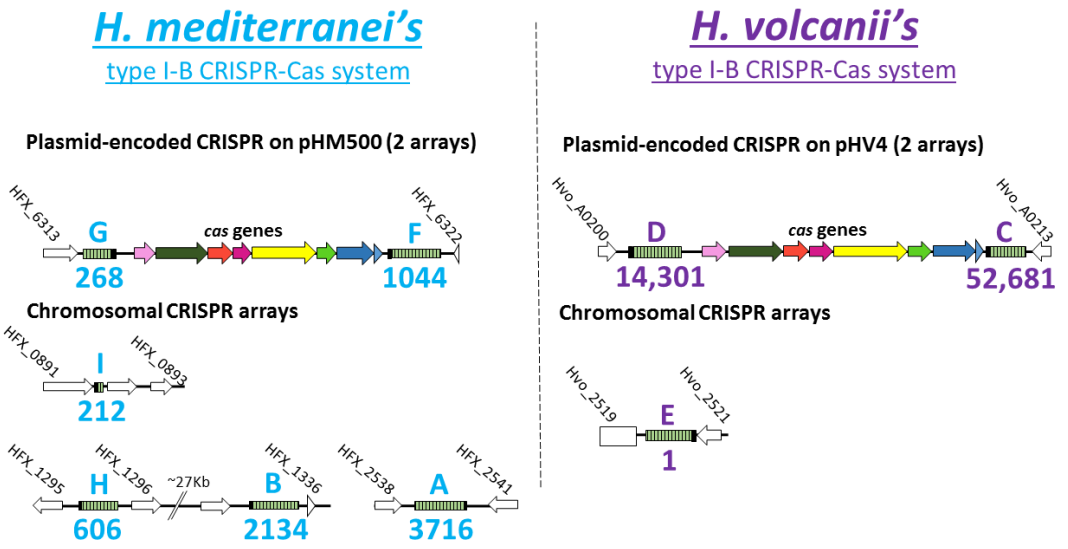


c.



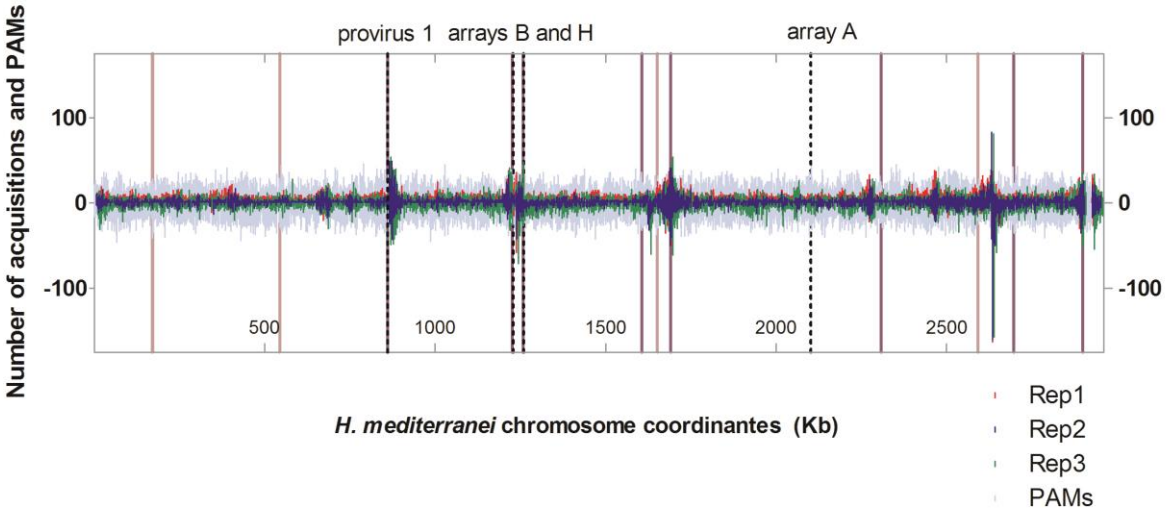
Supplementary Figure 3: Replicon representation in inter-species mating products sequenced using community sequencing. Three biologically independent experiments were performed a. Mean number of reads per kb across each of the *H.*

mediterranei replicons. b. Mean number of reads per kb across each of the *H. volcanii* replicons. c. Mean total reads per replicon normalized by replicon length. Error bars represent the standard error of the mean.

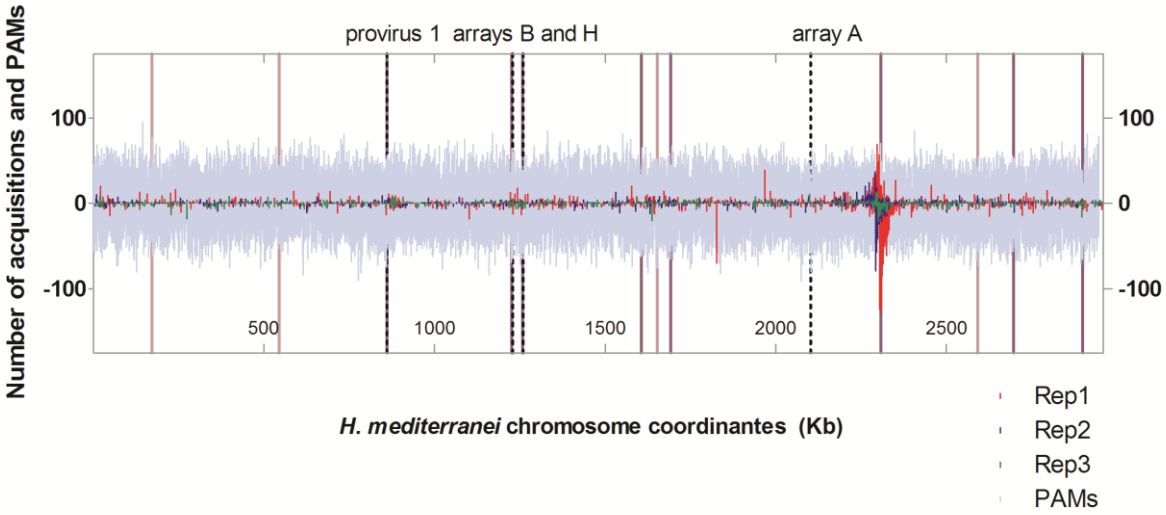


Supplementary Figure 4: Unique spacer acquisitions by array in each species following inter-species mating. Each CRISPR array of the two species is represented by a letter (A-I) with a number indicating the number of new unique spacers that were acquired in that aerray. Left, *H. mediterranei's* CRISPR-Cas system; *cas* genes are flanked by CRISPR arrays F and G on the pHM500 plasmid along with four additional chromosomal arrays A, B, H, and I. Right, *H. volcanii*'s CRISPR-Cas system; *cas* genes are flanked by CRISPR arrays C and D on the pHV4 plasmid along with additional chromosomal array E.

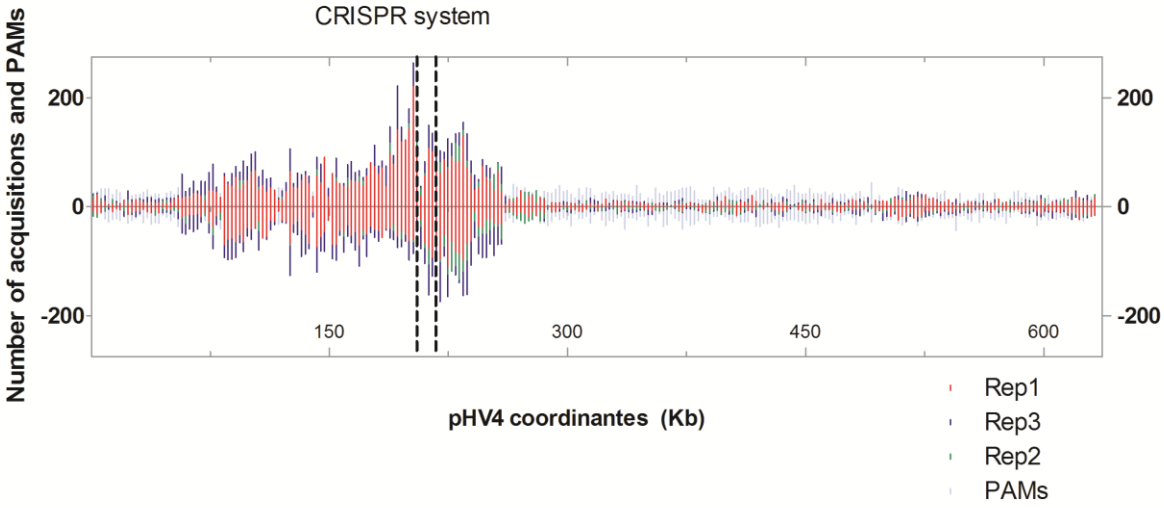
a.


provirus 1 arrays B and H
array A
Number of acquisitions and PAMs
100
0
-100
500
1000
1500
2000
2500
H. mediterranei chromosome coordinantes (Kb)
Rep1
Rep2
Rep3
PAMs

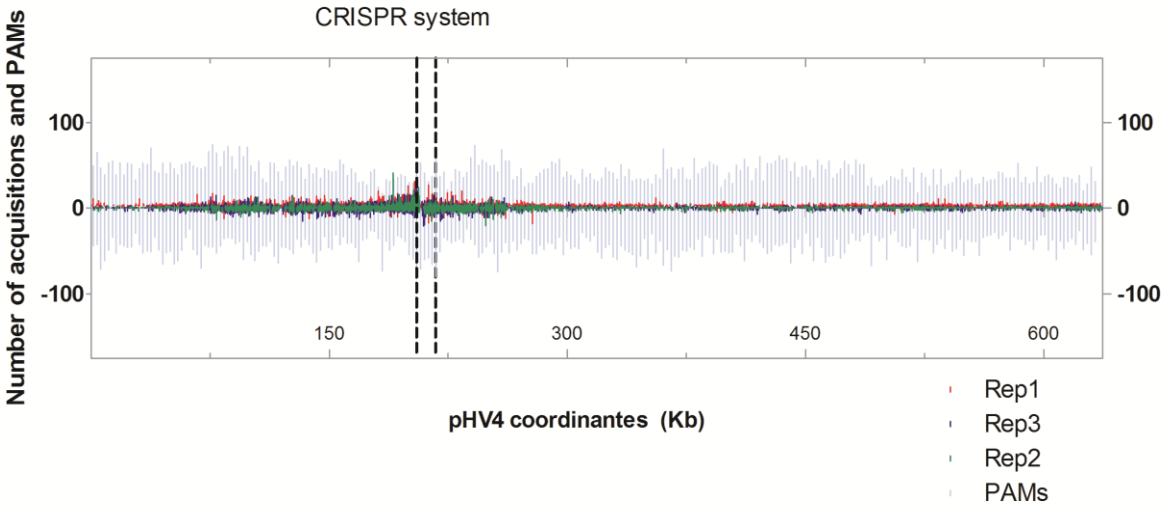
b.


provirus 1 arrays B and H
array A
Number of acquisitions and PAMs
100
0
-100
500
1000
1500
2000
2500
H. mediterranei chromosome coordinantes (Kb)
Rep1
Rep2
Rep3
PAMs

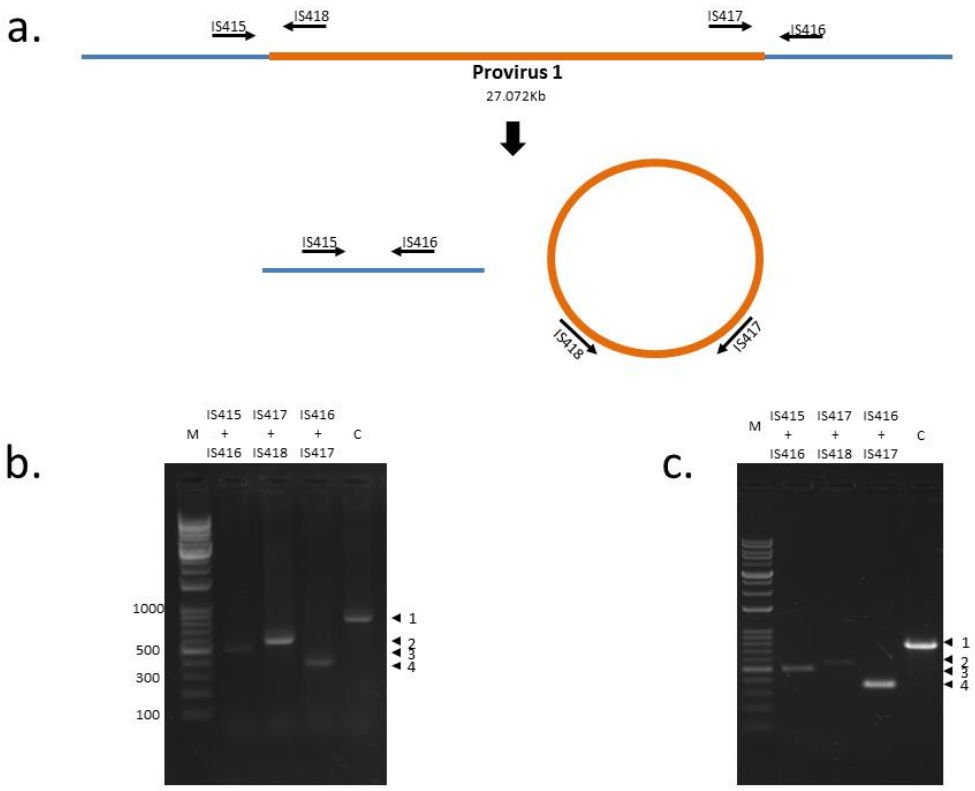
c.



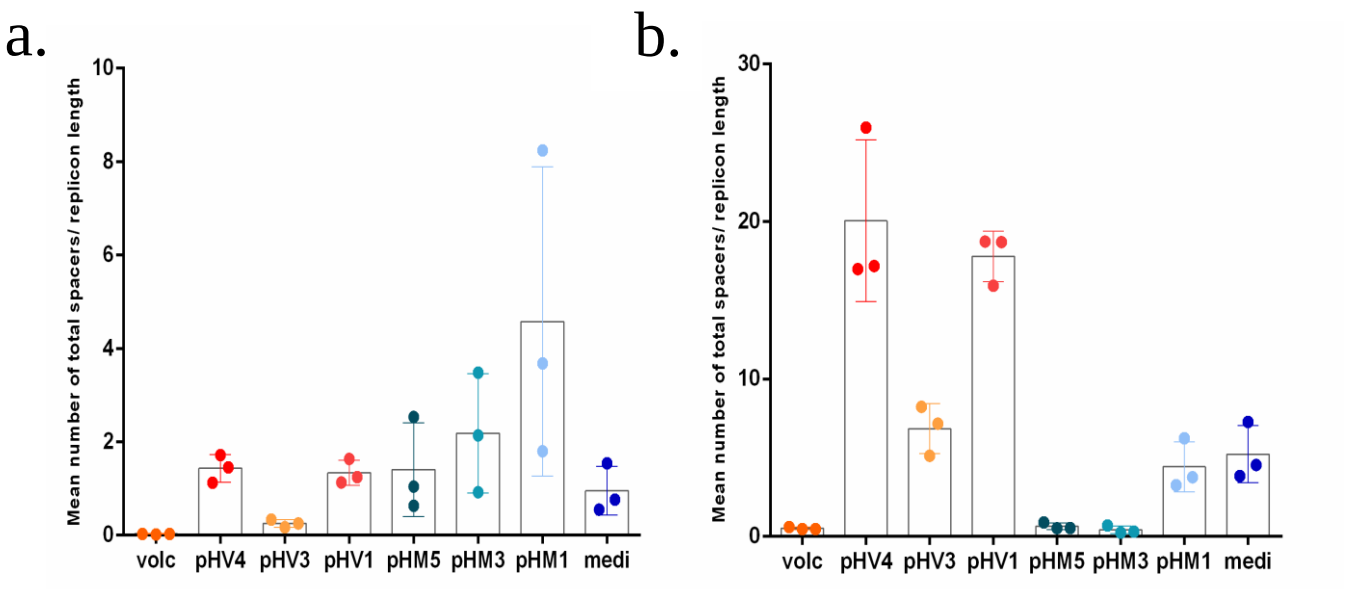
d.



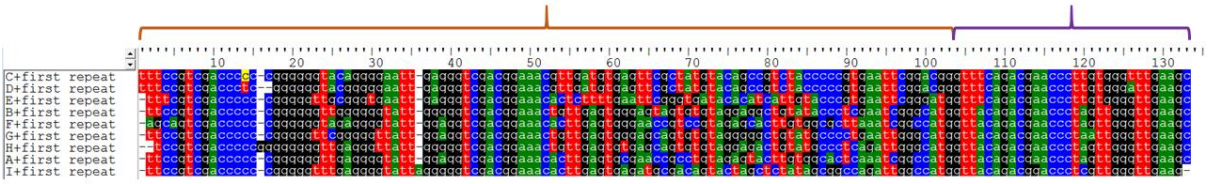
Supplementary Figure 5: Acquisition landscape following inter-species mating. Total spacers are colored by biological replicate and with PAMs shaded in grey. Dotted lines mark locations of CRISPR arrays and provirus 1. Purple lines mark locations of genes encoding integrases and recombinases and genes encoding transposases are denoted by pink lines. **a.** *H. volcanii* PAM signature and spacer acquisitions from *H. mediterranei* chromosome (2458 bp bins). **b.** *H. mediterranei* acquisitions from its own main chromosome (1160 bp bins). **c.** *H. volcanii* acquisitions from pHV4 (2435 bp bins). **d.** *H. mediterranei* acquisitions from pHV4 (2435 bp bins).



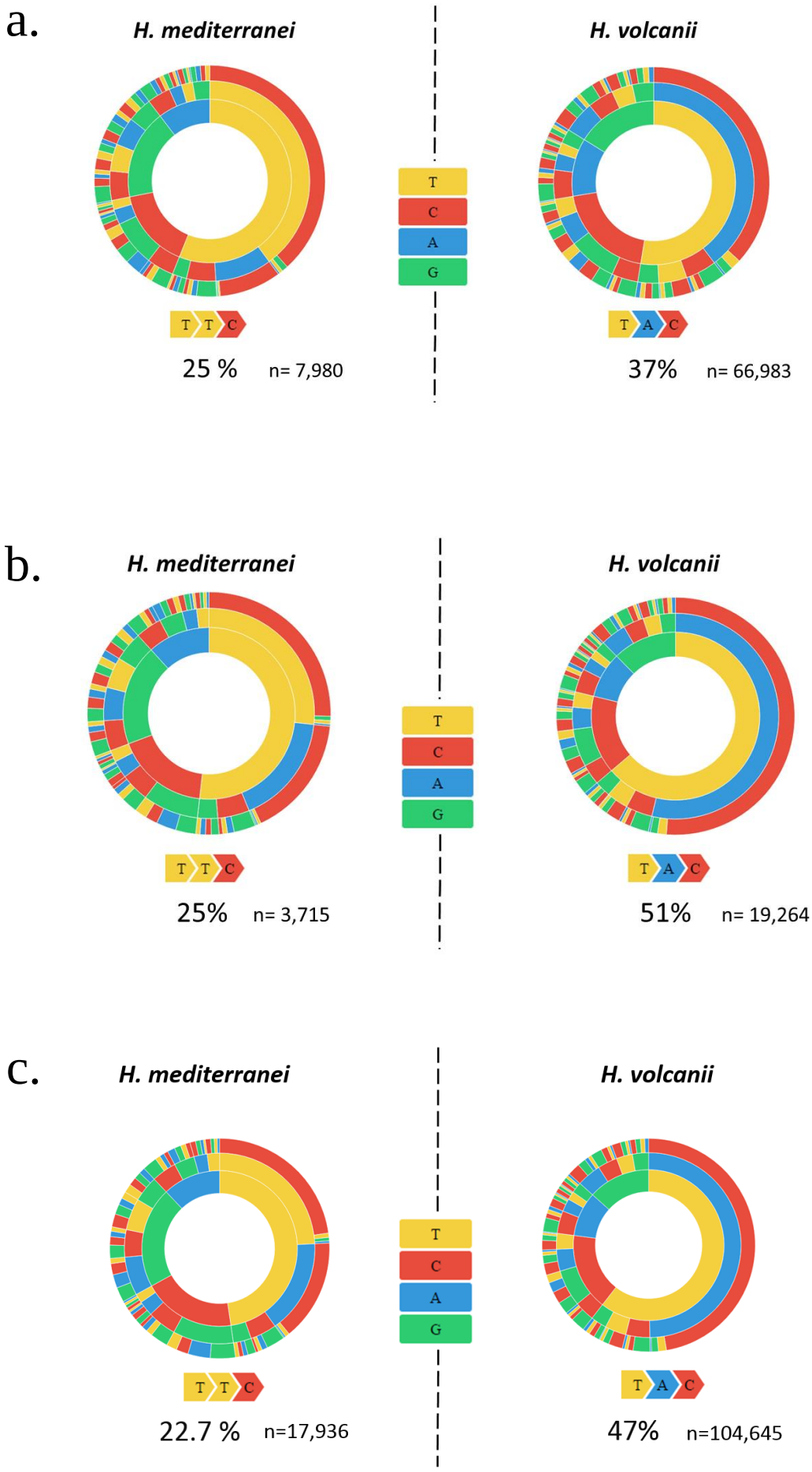
Supplementary Figure 6: Excision of provirus 1 of *H. mediterranei* from the chromosome is increased during inter-species mating compared to standard growth. a. Schematic representation of the primers location on provirus 1 locus in *H. mediterranei*. **b.** PCR results with primers from A where the templates were DNA extractions from *H. volcanii* and *H. mediterranei* mating products (the same DNA batch used for the spacer acquisition analysis), visualized by agarose gel electrophoresis. Similar results were obtained in two additional independent experiments. **c.** PCR results with the same sets of primers where the template was DNA extractions from a single *H. mediterranei* colony. Lanes: 1. *H. mediterranei* chromosomal *aceB* gene amplified with its own specific primers as a positive control, 2. The PCR amplicon indicating the presence of circular form of the provirus, 3. The PCR amplicon indicating the repaired chromosome following the excision of provirus 1, 4. The PCR amplicon indicating the presence of the integrated provirus in the chromosome. The ratio of the integrated form vs. the excised circular form based on gel densitometry was 16:1 when template was a standard colony of *H. mediterranei*, while that ratio shifted to 2.5:1 in the inter-species mating experiment. Similar results were obtained in one additional independent experiment.



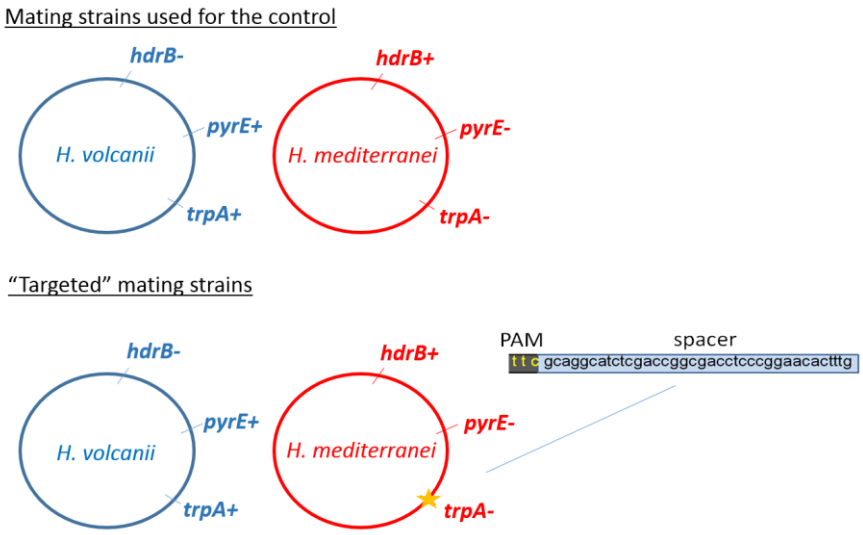
Supplementary Figure 7: Number of total spacers acquired during mating between *H. volcanii* and *H. mediterranei*. For each replicon, the mean of total spacers is shown normalized to replicon length. Error bars represent the standard error of the mean. Three biologically-independent experiments were performed **a**. Mean *H. mediterranei* spacer acquisitions. **b.** Mean *H. volcanii* spacer acquisitions.



Supplementary Figure 8: Sequence Alignment of the leader and first repeat. All 9 arrays leaders and first repeats from *H. volcanii* (C, D, E) and *H. mediterranei* (A, B, F, G, H, I) were aligned using ClustalX [2] as implemented in BioEdit. The leader end (left) is marked in brown and first repeat is marked in purple (right).


a.
H. mediterranei
H. volcanii
T
C
A
G
T T C
T A C
25 %
n= 7,980
37%
n= 66,983
b.
H. mediterranei
H. volcanii
T
C
A
G
T T C
T A C
25%
n= 3,715
51%
n= 19,264
c.
H. mediterranei
H. volcanii
T
C
A
G
T T C
T A C
22.7 %
n=17,936
47%
n=104,645

Supplementary Figure 9: Species-specific Protospacer Adjacent Motif (PAM) preference against all possible targets. The sequences of the three upstream bases from each individual protospacer were extracted. The relative abundances of these three-base PAMs were calculated, and are represented in a PAM wheel for each species. The favored PAM (appearing with highest frequency) for each species is marked. PAMs based on unique spacers (only one occurrence of each spacer is counted) either including. a. or excluding b. singleton spacers**.** c**.** PAM distribution when considering total spacer counts.



Supplementary Figure 10: Strains used for the mating efficiency experiments. The *H. mediterranei* "targeted" strain was engineered to contain a validated *H. volcanii* CRISPR spacer+PAM sequence, so that *H. volcanii* CRISPR-Cas could target it during mating. The location of the sequence insertion is marked with a star and corresponds to the locus of the deleted *trpA* gene.

Supplementary Tables

Supplementary Table 1: Haloarchaea with CRISPR arrays categorized by the presence of cross-targeting inter-species spacers. Haloarchaeal genomes from the database that had CRISPR arrays but did not have spacers that match other species were listed under "do not target other haloarchaea" and those which have at least one such spacer are listed as "target other haloarchaea".

Target other haloarchaea	**Do not target other haloarchaea**
Haloquadratum walsbyi DSM 16790	*Natronococcus occultus* SP4 **(only 1 spacer total)**
Haloferax volcanii DS2	*Haloarcula hispanica* ATCC 33960
Halomicrobium mukohataei DSM 12286	*Haloarcula hispanica* N601
Haloarcula marismortui ATCC 43049	*Halogeometricum borinquense* DSM 11551 **(only 5 spacers total)**
Haloquadratum walsbyi C23	*Natronomonas pharaonis* DSM 2160
Halorhabdus tiamatea SARL4B	*Natrialba magadii* ATCC 43099
Halorhabdus utahensis DSM 12940	*Natrinema pellirubrum* DSM 15624
Halorubrum lacusprofundi ATCC 49239	*Halostagnicola larsenii* XH-48 GCF_000517625 **(only 3 spacers total)**
Haloferax mediterranei ATCC 33500	*Haloferax gibbonsii* ARA6 GCF_001190965 **(only 3 spacers total)**
Natrinema sp. J7-2	
Natronobacterium gregoryi SP2	
halophilic archaeon DL31	

Supplementary Table 2: Haloarchaeal spacers that matched viruses and other haloarchaea. Spacers that matched viruses and other species of haloarchaea are presented in the following table along with description that includes the length of the spacer, location of the array (plasmid/chromosome), spacer sequence and target information (name of the species/virus, target gene, e value etc.) are provided.

Haloarchaea targeting	Length of spacer	Location of the array	Spacer sequence	Target
DL31	36 bp	chromosome	>spacer39 GCCGAGGAGCGACGTGCTCGTGACGGTCGCGGCGGC	*Haloferax volcanii* DS2, complete genome , npdG gene(F420H2:NADP oxidoreductase)score= 41.0 bits(44), e value, 8e-04 30/36(88%), stretch= 2-35
Natrialba magadii ATCC 43099	35 bp	chromosome	>>spacer4 TTCCGTGTCAACGTCGTCGTCATCGTCGTCATCGT	Pandoravirus inopinatum isolate KlaHel, complete genome,score= 44.6 bits(48), e value= 7e-04 24/24(100%), stretch=12-35
Haloquadratum walsbyi DSM 16790	36 bp 41 bp 42 bp	chromosome chromosome chromosome	>spacer1 GAAGTGTTCGCGGGGATGCGCCAAGACGCCATCCCGCTG >spacer3 GAAGTGCCGTCCCGGTGACTGTCACCGAGGGTCAAGAGATG >spacer4 CGAGTTCGCCGATAAGACGAGTGCTGCAAGTCGGAGTGACGG	*Haloquadratum walsbyi* C23 complete genome: CRISPR-associated exonuclease Cas4, cas4B, 3e-06 31/32(97%), spacerlength: 39 bp, score= 53.6 bits(58), strech: 8-39 *Haloquadratum walsbyi* C23 complete genome: Pseudo product: CRISPR-associated endoribonuclease Cas6 (nonfunctional)gene: cas6Bgene has a frameshift,1e-04 32/36(89%), spacerlength: 41 bp, score= 48.2 bits(52), strech: 6-41 *Haloarcula hispanica* N601 plasmid pHH126: cas3 helicase,1e-04 31/34(91%), spacerlength: 42 bp, score= 48.2 bits(52), strech:6-39
Haloferax volcanii	35 bp	plasmid pHV4	>spacer21 TATCACCCGTGGGCCGGTGACAACGAGGACCACGA	*Halorubrum lacusprofundi* ATCC 49239 chromosome 1(main chromosome), complete sequence , Hlac_0754 - hypotetical protein, 7e-06 33/36(92%), 36 bp, strech= 1-36, score=51.8 bits(56)
			>spacer2	*Haloarcula* sp. CBA1115 plasmid unnamed5,

Halomicrobium mukohataei DSM 12286	35 bp	chromosome	TCCTACGACCTCGTCGGCGTCAACGGCTGGCCCGA	complete sequence Gene: SG26_20405 ,Location: 16,746..17,825,e value= 8e-05 29/31(94%), score=48.2 bits(52), stretch=3-33, Archaeal BJ1 virus complete genome, hypothetical protein, Location: 27,949..29,031, e value=3e-04 30/33(91%), score=46.4 bits(50), strech=3-35
Haloarcula marismortui ATCC 43049	37 bp	plasmid pNG300	>spacer23 CGGCTGACTCCGGTGTGAGCGTGCACGCCCAGCCATC	*Haloarcula* sp. CBA1115, complete genome,hypothetical protein ,Qualifiers: Pseudo, Comment: disrupted, Location: complement(2,875,401..2,876,801), 3e-05 32/35(91%), stetch=3-37, score=50.0 bits(54)
	37 bp	(plasmid pNG300)	>spacer70 CGGCTGACTCCGGTGTGAGCGTGCACGCCCAGCCATC	*Haloarcula* sp. CBA1115, complete genome,Gene: hypothetical protein, Qualifiers: Pseudo, Comment: disrupted, Location: complement(2,875,401..2,876,801), e value=3e-05 32/35(91%) , score=50.0 bits(54) , stretch=3-37
	37 bp	(plasmid pNG400)	>spacer106 GAAGTCCACCGTATCACTGGACCGCGAAGTTCGGAAA	Halovirus HSTV-1, complete genome,CDS: AGC34548.1: portal , Portal (N- terminus) and Mu F-like (C-terminus), Location: 1,747..4,392 e value=4e-09 36/37(97%), stetch=1-37, score=62.6 bits(68),
	37 bp	(plasmid pNG400)	>spacer108 CCCCGCCGGTACTGCGGGGGCTCCAGCGAGTCGTCTA	Environmental Halophage eHP-31, partial genomeCDS: AFH22613.1: hypothetical protein, Qualifiers: Partial start, Title: hypothetical protein, Location: 1..1,285, 5e-08 32/32(100%), strech=4-35 , score= 59.0 bits(64)
	37 bp	(plasmid pNG400)	>spacer109 CCCCGCCGGTACTGCGGGGGCTCCAGCGAGTCGTCTA	Environmental Halophage eHP-31, partial genomeCDS: AFH22613.1: hypothetical protein, Qualifiers: Partial start, Title: hypothetical protein, Location: 1..1,285, e value=5e-08 32/32(100%), stretch=4-35, score=59.0 bits(64)

	36 bp	(plasmid pNG400)	>spacer111 CATGCACGACGGCGACGTGTCGGTGCTCAACTGCCT	Environmental Halophage eHP-14, partial genome, CDS: AFH21997.1: PCNA-like protein, Title: PCNA-like protein ,Comment: COG0592 DNA polymerase sliding clamp subunit (PCNA homolog); similar to Natrialba phage PhiCh1 PCNA, Location: complement(12,390..13,154), score=46.4 bits(50) , e value=3e-04 28/30(93%) , stratch=2-31
	36 bp	(plasmid pNG400)	>spacer125 CGGCTCTGAGTATGAGGGAGTCCTGTTCAGCCAAGA	Haloarcula hispanica N601 chromosome 1, complete sequence,CDS: AHB64487.1: hypothetical protein, Title: hypothetical protein, Location: 26,349..28,157, e value=8e-06 33/36(92%), stretch=1-36 , score= 51.8 bits(56)
	36 bp	(plasmid pNG400)	>spacer127 CTGCGGCGGCTGTTTGAGTTGTACCTTTCGCTGGTC	Halovirus HSTV-1, complete genome, CDS: AGC34590.1: MCM DNA helicase, Title: MCM DNA helicase, Location: 26,195..28,327, e value=1e-04 32/36(89%), score=48.2 bits(52), stretch= 1-36
	37 bp	(plasmid pNG400)	>spacer130 GCCGCAACGAAGCCCGGCAGGAACTGGGGCGTGACGA	Halovirus HSTV-1, complete genome,CDS: AGC34548.1: portal, Title: portal, Comment: Portal (N- terminus) and Mu F-like (C-terminus), Location: 1,747..4,392, e value=4e-09 36/37(97%), score=62.6 bits(68), stretch=1-37
	37 bp	(plasmid pNG400)	>spacer132 GCGTTGGTCGAGTTTGACCGTGCCGTAGATTGTCGCG	Halovirus HSTV-1, complete genome, CDS: AGC34548.1: portal, Title: portal, Comment: Portal (N- terminus) and Mu F-like (C-terminus),Location: 1,747..4,392, e value=2e-06 34/37(92%), score=53.6 bits(58), stretch=1-37
	36 bp	(plasmid pNG400)	>spacer136 GGCCGGCACGGCGTCACGCGAGAGATGTGTCCTTGC	Halovirus HSTV-1, complete genome, CDS: AGC34590.1: MCM DNA helicase, Title:

	37 bp 36 bp 36 bp	(plasmid pNG400) (plasmid pNG400) (plasmid pNG400)	>spacer146 GTACTCGTCACACCACGGGCACACGCCGTCGTCAATC >spacer147 CAGTGGGACGTGTTCGGCCATGCCGTAGATAGACAG >spacer170 AGCACATAGTCGATGCTCGGGTCGTGCCAGTGGCCG	MCM DNA helicase, Location: 26,195..28,327, e value=8e-06 33/36(92%), score=51.8 bits(56), stretch=1-36 Halovirus HSTV-1, complete genome, CDS: AGC34560.1: hypothetical protein, Title: hypothetical protein, Location: 10,838..11,209, e value=2e-06 32/34(94%), score=53.6 bits(58) stretch=4-37 Halovirus HSTV-1, complete genome, CDS: AGC34558.1: head maturation protease, Title: head maturation protease, Location: 8,151..9,572, score=55.4 bits(60), stretch=2-36, e value= 6e-07 33/35(94%) Environmental Halophage eHP-14, partial genome,CDS: AFH21988.1: YonJ-like protein, Title: YonJ-like protein, Comment: COG1311 archaeal DNA polymerase II, small subunit/DNA polymerase delta, subunit B, Location: complement(3,763..4,923), e value=8e-06 33/36(92%), score=51.8 bits(56), stretch=1-36
Haloquadratum walsbyi C23	35 bp 34 bp 36 bp 33 bp	chromosome chromosome chromosome chromosome	>spacer2 AAACCGAGTGACTGCCAGGCGTCTCCTGCCTGCCG >spacer10 CACGTTTGGTTGTCGGGCTGGCACGGTCAACCGC >spacer19 TGCACGGAACTCTCGAATATCGATAAGATTACCTGT >spacer20 ATGACTCCGACAGAGGACCCAGTGAATACTGTT	Uncultured virus contig036 genomic sequence,e value=5e-07 33/35(94%), score=55.4 bits(60), stretch= 1-35 Uncultured virus contig195 genomic sequence, e value=0.001 30/34(88%), score=44.6 bits(48), stretch=1-34 Environmental Halophage eHP-40, partial genome, e value=3e-10 36/36(100%), score=66.2 bits(72), stretch=1-36 Environmental Halophage eHP-40, partial genome, e value=6e-06 31/33(94%), score=51.8 bits(56), stretch=1-33

	37 bp	chromosome	>spacer21 ACATGGGCAACTGGCGTAGGGTTGACTGCCGTAGAGG	Environmental Halophage eHP-5, partial genome, e value=2e-06 34/37(92%), score=53.6 bits(58), stretch=1-37
	36 bp	chromosome	>spacer25 TTTCAAAGCCGACGGCGAACAGCACTGCCGGCAAGT	Uncultured virus contig026 genomic sequence, e value=3e-10 36/36(100%), score=66.2 bits(72), stretch=1-36
	35 bp	chromosome	>spacer29 AGCACGGCTAACGACGTGGTGGATCTAAGTAAGCC	Uncultured virus contig003 genomic sequence, e value= 0.001 27/29(93%), score=44.6 bits(48), stretch=1-29
	35 bp	chromosome	>spacer34 GATGATCAGCCAGTTCAACATGATGATCCAGGTCG	Uncultured virus contig003 genomic sequence, e value=8e-05 30/35(86%), score=48.2 bits(52), stretch= 1-35
	34 bp	chromosome	>spacer50 CGTTGAGATGTCGGTCGCCTGTTTTTGTGCGCTG	Uncultured virus contig015 genomic sequence, e value=8e-05 31/34(91%), score=48.2 bits(52), stretch=1-34
	35 bp	chromosome	>spacer56 CGATATTAACGTTTGAATCAAACGAATAATCAATA	Environmental Halophage eHP-40, partial genome, e value=2e-05 32/35(91%), score=50.0 bits(54), stretch=1-35
	36 bp	chromosome	>spacer65 AAGCTCGAAGGTGCGCGTTGGTGATGCACCGAACTC	Uncultured virus contig015 genomic sequence, e value=3e-09 35/36(97%), score=62.6 bits(68), stretch=1-36
	34 bp	chromosome	>spacer69 AGGCAGTCACACCGCCGATGACGGGAGTGCAGAC	Uncultured virus contig015 genomic sequence, e value=5e-07 30/30(100%), score=55.4 bits(60), stretch=5-34
	36 bp	chromosome	>spacer70 GACCTGAGATATGATGATGACCCATAAGACAAACAT	Environmental Halophage eHP-2, partial genome 1e-08 35/36(97%) , 60.8 bits(66)
	37 bp	chromosome	>spacer78 ACACTCGGCACAAAATGCGGTCAGAAACAGACCGCCG	Uncultured virus contig015 genomic sequence, 8e-11 37/37(100%), 68.0 bits(74)
			>spacer79	

	36 bp	chromosome	GGCTTGTGCTGCGTTCTCGGCTGTTTTGCGTGGTGT	Uncultured virus contig003 genomic sequence,6e-06 33/36(92%) , 51.8 bits(56)
			>spacer84	
			ATTTTCGTTTAGGGAGGGATCCTGGTGTCGTAGGA	*Haloquadratum walsbyi* DSM 16790 complete genome, CDS: CAJ52277.1: uncharacterized protein, Title: uncharacterized protein, Location: complement(1,384,314..1,385,258), e value=1e-09 35/35(100%), score=64.4 bits(70), stretch=1-35
	35 bp	chromosome		
Halorhabdus tiamatea SAL4B			>spacer51	
	37 bp	chromosome	GCTAGACAGCAGTCGATTTGGGCTATTTTGGCTAGTA	*Halobacterium* sp. DL1, complete genome, Gene: ATP-binding protein, Qualifiers: Pseudo, Comment: disrupted, Location: 515,496..517,610, score=62.6 bits(68) , e value=4e-09 36/37(97%) ,sretch=1-37
Halorhabdus utahensis DSM 12940			>spacer21	
	35 bp	chromosome	TCGGTCTCGACGCCGACGAGTACCGCGAGCAACTC	*Halopenitus persicus* DNA, complete genome, strain: CBA1233, blast the gene and resulted: NAD-dependent; catalyzes the oxidative decarboxylation of malate to form pyruvate; does not decarboxylate oxaloacetate, score=44.6 bits(48), e value 0.001 24/24(100%), stretch-7-30
			>spacer25	
	34 bp	chromosome	CCAGCATGGAGGCCGGCCCGTCGAACTACCCGAC	icosahedral virus 1, complete genome (Haloarcula californiae),CDS: ALJ99672.1: hypothetical protein, Title: hypothetical protein, Comment: VP1; gene 9, Location: 3,472..8,298,score=51.8 bits(56) ,e value= 6e-06 28/28(100%) , stretch=6-33
			>spacer35	
	37 bp	chromosome	CCGTCGACGCGATCGTCGCCCAGACCGAACCGATCGA	*Natronobacterium gregoryi* SP2, complete genome,Gene: Natgr_3439, Comment: IMG reference gene:2510573760, Location: 3,400,991..3,401,836, score=53.6 bits(58) , e value= 2e-06 34/37(92%), stretch= 1-37
			>spacer42	
	36 bp	chromosome	GCCCGAGGACGTCTACCGCGAGCACATCCTCGCGAC	*Halorubrum lacusprofundi* ATCC 49239 chromosome 1, complete sequence, Gene:

	37 bp	chromosome	>spacer43 CATCGCCCGCCTTCTGGTCGACCGCGCGGAGAATCTC	Hlac_0755, Location: 759,925..761,565,CDS: ACM56355.1: hypothetical protein, Title: hypothetical protein, Comment: KEGG: gtn:GTNG_2829 terminase large subunit, putative,score=48.2 bits(52) , e value=8e-05 32/36(89%), stretch=1-36 *Halorubrum lacusprofundi* ATCC 49239 chromosome 1, complete sequence, Gene: Hlac_0759, Location: 764,530..765,537, Length: 1,008, CDS: ACM56359.1: hypothetical protein, Title: hypothetical protein, Comment: KEGG: LOC689900; hypothetical protein LOC689900, score= 46.4 bits(50), e value= 3e-04 31/35(89%), stretch= 3-37
Halorubrum lacusprofundi ATCC 49239	35 bp 36 bp 32 bp	plasmid pHLAC01 plasmid pHLAC01 plasmid pHLAC01	>spacer16 GGGACGGTCTACGTCGGCAGCGACGATAACAGCCT >spacer90 CTGTGGAACCTTAAACGAAGTCTCGAACGCACGCTC >spacer105 GCGACGGTCGAGAAGAGCGATTCCTAGATAGT	*Halorhabdus utahensis* DSM 12940, complete genome, ene: Huta_1990, Location: 2,012,531..2,013,922, Length: 1,392, CDS: ACV12158.1: Pyrrolo-quinoline quinone, Title: Pyrrolo-quinoline quinone, Comment: PFAM: Pyrrolo-quinoline quinone~SMART: Pyrrolo-quinoline quinone~KEGG: nph:NP5062A cell surface protein/lipoprotein, score=44.6 bits(48) e value=0.001 26/27(96%), stretch=1-27 *Haloterrigena turkmenica* DSM 5511, complete genomeGene: Htur_1455, Location: complement(1,509,258..1,513,622), CDS: ADB60341.1: Pyrrolo-quinoline quinone, Title: Pyrrolo-quinoline quinone *Halobacterium* sp. DL1 plasmid, complete sequence, CDS: AHG05589.1: hypothetical protein, Title: hypothetical protein, Location: complement(214,445..215,200), score=66.2 bits(72), e value=3e-10 36/36(100%), stretch=1-36 *Natrialba magadii* ATCC 43099, complete genome, Gene: Nmag_2811, Location: 2,885,847..2,886,548CDS: ADD06365.1:

				Haloacid dehalogenase domain protein hydrolase, Title: Haloacid, dehalogenase domain protein hydrolase, Comment: PFAM: Haloacid,dehalogenase domain protein hydrolase~KEGG: hut:Huta_1147 HAD-superfamily hydrolase, subfamily IA, variant 1 , score=46.4 bits(50) 3e-04 27/28(96%), stretch=1-28
			>spacer130	*Halobacterium* sp. NRC-1, complete genome, CDS: AAG19574.1: glutamate dehydrogenase, Title: glutamate dehydrogenase, Location: complement(903,818..905,068), score=44.6 bits(48) 0.001 27/29(93%), stretch=1-29
				Halobacterium salinarum R1 complete genome,CDS: CAP13852.1: glutamate dehydrogenase, Title: glutamate dehydrogenase, Location: complement(895,807..897,057), score=44.6 bits(48), e value-0.001 27/29(93%), stretch=1-29
	34 bp	plasmid pHLAC01	CGGGATCGCGGACGTCATCACCGTCGACGAAGGC	
			>spacer134	*Halorubrum trapanicum* DNA, complete genome, strain: CBA1232, from blastn the targeted protein is Title: orc1/cdc6 family replication initiation protein(from Halorubrum lacusprofundi), score=66.2 bits(72), e value= 3e-10 36/36(100%)' stretch 1-36
				also many other haloarchea and also self* : Halorubrum lacusprofundi ATCC 49239 chromosome 1, complete sequence, orc1/cdc6 family replication initiation protein, score=57.2 bits(62), e value=1e-08 34/36(94%), stretch=1-36
	36 bp	plasmid pHLAC01	CACGAGCGCGTGGTCGTGATCATGCTCGACGAGATC	
			>spacer141	
				Haloferax gibbonsii strain ARA6, complete genome,Gene: ABY42_00845, Location: complement(159,515..160,258), Title: DNA polymerase sliding clamp, score=60.8 bits(66), e value= 1e-08 39/43(91%), stretch=1-43
	43 bp	plasmid pHLAC01	ACGAACAGACACGCAAGCTCAACATCCACATCGACGGGCTGTC	
			>spacer150	
	36 bp	plasmid pHLAC01	AGAACCGCGCGTCGGGGACGTCGAAACAGGCGAACG	Halorubrum phage CGphi46 genomic

	36 bp (same seq as spacer 150) 33 bp	plasmid pHLAC01 plasmid pHLAC01	>>spacer150 AGAACCGCGCGTCGGGGACGTCGAAACAGGCGAACG >spacer192 GAAGCAGGGGCTTGCCCCTGACCCGACCGGCGG	sequence, CDS: AGN33798.1: hypothetical protein, Title: hypothetical protein, Location: complement(8,550..10,031), score=48.2 bits(52), e value=8e-05 32/36(89%), stretch=1-36 *Halorubrum* phage CGphi46 genomic sequence, CDS: AGN33798.1: hypothetical protein, Title: hypothetical protein, Location: complement(8,550..10,031), score=48.2 bits(52), e value=8e-05 32/36(89%), stretch=1-36 *Haloferax volcanii* DS2, complete genome, intergenic region between HVO_1434 (Title: homolog to HRPV1-ORF1 ,Comment: identified by glimmer; putative) and HVO_3034(tRNA-Arg), score=46.4 bits(50) e value=3e-04 25/25(100%), stretch=1-25.
Haloferax mediterranei ATCC 33500	35 bp 38 bp	chromosome chromosome	>spacer5 GTGCATCGCCGAGCGGTCGTTCGGGTCCATCTTGT >spacer55 GACGCTCGAAGCCGGCGCGCTCGTCCTCGCCGACAAAG	*Salinarchaeum* sp. Harcht-Bsk1, complete genome, CDS: AGN02478.1: MCM family protein, Title: MCM family protein, Comment: COG1241 Predicted ATPase involved in replication, control, Cdc46/Mcm family, Location: complement(2,755,120..2,757,213), score=46.4 bits(50), e value= 8e-04 31/35(89%), stretch=1-35 *Halopiger xanaduensis* SH-6, complete genom, CDS: AEH36243.1: MCM family protein, Title: MCM family protein,, Comment: KEGG: htu:Htur_0316 MCM family protein~PFAM: DNA-dependent ATPase MCM~SMART: DNA-dependent ATPase MCM; ATPase, AAA+ type, core, Location: complement(923,814..925,922), score= 46.4 bits(50) , e value =8e-04 31/35(89%), stretch=1-35 *Haloferax gibbonsii* strain ARA6, complete genome, score=60.8 bits(66), e value= 4e-08 35/36(97%), stretch=1-36 *Haloferax volcanii* DS2, complete genome,

	37 bp	plasmid phm500	>spacer76 CTTTGTCGGCGAGGACGAGCGCGCCGGCTTCGAGCGT	Title: ATP-dependent DNA helicase MCM, Location: 199,368..201,476, score=60.8 bits(66)e value= 4e-08 35/36(97%), stretch=1-36 *Haloferax gibbonsii* strain ARA6, complete genome, score=60.8 bits(66), e value= 4e-08 35/36(97%), stretch=1-36 Haloferax volcanii DS2, complete genome, Title: ATP-dependent DNA helicase MCM, Location: 199,368..201,476, score=60.8 bits(66)e value= 4e-08 35/36(97%), stretch=1-36
Natrinema sp. J7-2	36 bp 35 bp	chromosome chromosome	>spacer58 ATGAATATGGGGACTCGAGAGACTCACCCTGATCGC >spacer59 GCCGACGAGTTCGAGGCGTTCGTCGACGCCGGCGA	*Natrinema* virus SNJ1, complete genome, score=48.2 bits(52), e value= 9e-05 29/31(94%), stretch=1-31 *Natronobacterium gregoryi* SP2, complete genome, CDS: AFZ74554.1: hypothetical protein, Title: hypothetical protein, Location: 3,397,627..3,399,306 (probobly capsid protein from blastp), score=55.4 bits(60) , e value= 6e-07 33/35(94%), stretch=1-35
Natronobacterium gregoryi SP2	34 bp 36 bp	chromosome chromosome	>spacer1 TTGTCGTGTACGCGTGTGTTCGACGAGTTCGAGT >spacer41 CCACCACTTGTTCGCCTTTCGCGTGCGTGTGCCCGT	*Natrinema pellirubrum* DSM 15624, complete genome, Comment: PFAM: Transposase DDE domain, Location: 266,076..267,299, score=57.2 bits(62), e value= 1e-07 33/34(97%), stretch=1-34 Environmental Halophage eHP-18, partial genome, score=48.2 bits(52) 7e-05 32/36(89%), strerch=1-36

* *Halorubrum lacusprofundi* , was the only strain that harbored a spacer that had a significant match to its own genome (matching the *orc1/cdc6* gene). However, this spacer had two mismatches, indicating that it is more likely to have been acquired from a different species rather than being a case of CRISPR autoimmunity, thought to be rare in archaea [3].

Supplementary Table 3: Cross-targeting spacers of environmental isolates from Atlit. Spacers that matched other species of isolated haloarchaea from Atlit with high identity are presented in the following table. All cross-targeting spacers from each isolate are listed with description regarding the spacers sequence, the name of the targeted isolate, the length of the targeted contig sequence and the inference of the nature of the target (plasmid/chromosome). Identical spacers detected in different isolates are marked in blue.

Isolate name	Spacer seq	Isolate targeted	Match contig length(Kb)	Putative replicon
7R (*Haloarcula*)	spacer3: AAATCTCGCCGACAGCACCGCTCACATCCCGGCTG	self, 47R, 120R	288.3;225.6	Plasmid
	spacer15:TGCTGACTCCGGAGTGAGTGTGCACGCCCAGCCATC	47R,120R	225.6	Plasmid
	spacer53: CATCATGAGTTCACGAGACCGCCGGGTACAGGTCGCCC	120R	1393	Chromosome
	spacer76:CTTCGAGGTTCAGCGAGTCCGACCCGCCAGCCGATA	47R,120R	1.25	Plasmid
	spacer77:TCACCCACCCAGCAAGTGCGGCAACGGCAACGGCGA	47R,120R	43	Plasmid
	spacer80: ATCGAGGACGTGCTGGACCGCCACGGCGACTGGTG	47R,120R	7.5	Plasmid
	spacer 81: GCTGGCGACGAGTCGCTGGTCGACGCGCTGGGCATCAACA	120R	7.3	Plasmid
19N (*Haloferax*)	spacer 164: ATGCCGACGACGAGGAGCGCGGTCGTCCAGCCGACC	47N, 31R	1,663	Chromosome
24N (*Haloferax*)	spacer 4 : TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG	19N, 48N, 120R	203	Chromosome
	spacer34: CTTGATGCGGTCGTTCATCCGGTTGAGACACCGCAGA	self, 19N, 48N, 47N	277.5	Plasmid
	spacer3: AAAGACGTACTCTCGCTCGGCGGTCCAGTTCGGCGAC	19N, 48N, 120R	203	Chromosome
	spacer 54: ACTGAACGAGTCGCCGTTGACTTCCTGAGCGACCTGCC	19N, 48N	203	Chromosome
48N (*Haloferax*)	spacer 6: TACGTCAACGACTCCGATGCACCTACAACAGCCCAGA	47N	1663	Chromosome
47N (*Haloferax*)	spacer 5:ATGGTCAACCTCCTCTGCAACGGGCTCATGACCGC	19N, 48N	203	Chromosome
	spacer 17: TCGCCCGCGGACGCAATCGACGCCGGAATCGGTATGATT	self, 48N, 24N	1663	Chromosome
	spacer 52: GTTGATGACGACAGCGCAGGACTTGGACTCGACTTCAG	48N	203	Chromosome
	spacer 53: TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG	19N, 48N	203	Chromosome

Supplementary Table 4: Perfect match cross-targeting spacers from the environmental isolates. Spacers that matched perfectly other species of isolated haloarchaea from Atlit are presented in the following table. All cross-targeting spacers from each isolate are listed with the following description: the number of spacers in the array that the spacer is found on (column "spacers"), corresponding PAM sequence, origin array type and orientation (using CRISPRDetect [4]), location of the spacer in the array, gene target annotation and the spacers sequence. Identical spacers detected in different isolates are marked in blue.

Isolate Name	Arrays	Spacers	Targeting	PAM	Array type/orientation	Location in the array/identity of array (contig#)	Annotation	spacers sequence
7R_*haloarcula*	2	111	1.(spacer3) 47R, 120R, self 2.(spacer15) 47R, 120R 3.(spacer76) 47R 4.(spacer80) 47R, 120R 5.(spacer77) 47R, 120R	1.(spacer3) ccg 2.(spacer15) ttc 3.(spacer76) ttc 4.(spacer80) ttc 5.(spacer77) ttc	1.(spacer3) type I-B / forward 2.(spacer15) type I-B/ forward 3.(spacer76) type I-B/ forward 4.(spacer80) type I-B/ forward 5. (spacer77) type I-B/ forward	1.(spacer3) spacer 1out of 58 /contig2 2.(spacer15) spacer 13 out of 58 / contig 2 3.(spacer76) spacer12 out of 53 / contig 18 4.(spacer80) spacer 16 out of 53 / contig 18 5. (spacer77) spacer 13 out of 53 / contig 18	1.(spacer3) carbohydrate ABC transporter substrate-binding protein, CUT1 family (htr) 2.(spacer15) inter genic: btween repH2 and pNG3030 [*Haloarcula marismortui* ATCC 43049 plasmid pNG300] 3.(spacer76) hypothetical protein: Pseudo: disrupted [*Haloarcula* sp. CBA1115] 4.(spacer80) pNG3051 [*Haloarcula marismortui* ATCC 43049 plasmid pNG300] 5.(spacer77) pNG3048 [*Haloarcula marismortui* ATCC 43049 plasmid pNG300]	1.(spacer3) AAATCTCGCCGACAGCACCGCTCACATCCCGGCTG 2.(spacer15) TGCTGACTCCGGAGTGAGTGTGCACGCCCAGCCATC 3.(spacer76) CTTCGAGGTTCAGCGAGTCCGACCCGCCAGCCGATA 4.(spacer80) ATCGAGGACGTGCTGGACCGCCACGGCGACTGGTG 5.(spacer77) TCACCCACCCAGCAAGTGCGGCAACGGCAACGGCGA
19N_haloferax	5	154	1. (spacer 164) 47N	1. (spacer 164) gtg	1. (spacer 164) (not type I-B nor I-D)/ forward	1. (spacer 164) spacer 50 out of 50 /contig 3	1. (spacer 164) membrane protein [*Haloferax gibbonsii* strain ARA6]	1. (spacer 164) ATGCCGACGACGAGGAGCGCGGTCGTCCAGCCGACC

24N_*haloferax*	4	109	1. (spacer34) 47N, self 2. (spacer3) 48N,120R 3. (spacer4) 48N, 120R 4. (spacer54) 48N	1. (spacer34) ggc 2. (spacer3) ttc 3. (spacer4) ttc 4. (spacer54) ttc	1. (spacer34) type I-B/ forward 2. (spacer3) type I-B/ forward 3. (spacer4) type I-B/ forward 4. (spacer54) type I-B/ Reverse	1. (spacer34) spacer 32 out of 37 /contig2 2. (spacer3) spacer 1 out of 37 / contig 2 3.(spacer4) spacer 2 out of 37 / contig 2 4. (spacer54) spacer 33 out of 43 / contig 4	1. (spacer34) pstB2, ABC-type transport system ATP-binding protein (probable substrate phosphate) [*Haloferax volcanii* DS2 plasmid pHV4] 2. (spacer3)SG26_14165 long-chain fatty acid--CoA ligase [*Haloarcula* sp. CBA1115, complete genome] 3. (spacer4) no good hit to anything (not in haloarchaea)- in our isolates its hypotetical protein 4. (spacer54) tsgA3 ABC-type transport system periplasmic substrate-binding protein (probable substrate sugar) [*Haloferax volcanii* DS2, complete genome]	1. (spacer34) CTTGATGCGGTCGTTCATCCGGTTGAGACACCGCAGA 2. (spacer3) AAAGACGTACTCTCGCTCGGCGGTCCAGTTCGGCGAC 3. (spacer4) TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG 4. (spacer54) ACTGAACGAGTCGCCGTTGACTTCCTGAGCGACCTGCC
47N_*haloferax*	3	70	1. (spacer5) 48N 2. (spacer17) 48N, 24N, self 3. (spacer52) 48N* 4. (spacer53) 48N, 120R	1. (spacer5) atc 2. (spacer17) gat 3. (spacer52) ttc 4. (spacer53) ttc	1. (spacer5) type I-B/ forward 2. (spacer17) type I-B/ forward 3. (spacer52) type I-B/ forward 4. (spacer53) type I-B/ forward	1. (spacer5) spacer 3 out of 33 / contig 4 2. (spacer17) spacer 15 out of 33 / contig 4 3. (spacer52) spacer 1out of 32 / contig 5 4. (spacer53) spacer 2 out of 32 / contig 5	1. (spacer5) mrpD3- Mrp-type sodium/proton antiporter system subunit D3[*Natronomonas moolapensis*] 2. (spacer17) tsgD11-ABC-type transport system ATP-binding protein (probable substrate sugar) [*Haloferax volcanii* DS2] 3. (spacer52) Nmag_1905- alpha/beta hydrolase fold protein[*Natrialba magadii* ATCC 43099] 4. (spacer53) no good hit to anything (not in haloarchaea)- in our isolates its hypotetical protein	1. (spacer5) ATGGTCAACCTCCTCTGCAACGGGCTCATGACCGC 2. (spacer17) TCGCCCGCGGACGCAATCGACGCCGGAATCGGTATGATT 3. (spacer52) GTTGATGACGACAGCGCAGGACTTGGACTCGACTTCAG 4. (spacer53) TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG

Supplementary Table 5: List of spacers from the natural isolates that target viruses. Spacers that matched halovirus sequences from the databases with high identity are presented in the following table. All spacers targeting halo-viruses are listed with the following description: the isolate that had the spacer, spacer sequence, location of the array and the spacer in the array, the corresponding PAM sequence, spacer identity to the target, type of the array and orientation (using CRISPRDetect [4]) and the halovirus the spacer targets. Identical spacers detected in different isolates are marked in blue.

Isolate name	Spacer sequence	Location in the array/identity of array (contig#)	PAM	Identity spacer/target	Array type /orientation	Targeting
47N	TGGGTCGAACTCGCAAATGACG GCTCCACGTTCACGC	spacer 8 out of 8 /contig 5	caa	31/37	type I-B / forward	Halovirus HRTV-4, complete genome
47R	CTACCGGAGTGTGTCGAGGCGT GCATCCACGA	spacer 60 out of 68 /contig 8	gcg	30/32	NA / reverse	Halovirus HSTV-1, complete genome
	GACTTTGCGGAGTACCGACATA AAGGCAATGACCGCG	spacer 17 out of 26 /contig 11	tat	34/37	type I-B / forward	Haloarcula hispanica pleomorphic virus 1
	TGTGAGCGGTGACGTGCCGATA CCTGAGTCTGCTGC	spacer 21 out of 26 /contig 11	tac	33/36	type I-B / forward	Haloarcula hispanica pleomorphic virus 2
48N	GCGTGAACGTGGAGCCGTCATT TGCGAGTTCGACCCA	spacer 45 out of 45 /contig 4	caa	31/37	type I-B / reverse	Halovirus HRTV-4, complete genome
	ACGCGCCATCCGGACCATCGCC GCCGACGAGGCTGC	spacer 44 out of 45 /contig 4	ttc	30/36	type I-B / reverse	Archaeal BJ1 virus complete genome
120R	GATCTGTTCTAACTCTTCCTGC AGCTTCGACACATC	spacer 7 out of 21 /contig 9	gct	35/36	NA / reverse	*Haloarcula hispanica* pleomorphic virus 1
	CGGGAGTACTCCACGCCCTCGC GGTAGAACTCGTCT	spacer 7 out of 8 /contig 9	gcg	30/36	NA / forward	Halovirus HSTV-1, complete genome
26R	GTCTGGATTCGGACGCGCGACG GCGACGTGACGCG	spacer 13 out of 18 /contig 8	tac	34/35	type I-B / forward	Uncultured virus clone TS-May-2009-R-Contig-28 genomic sequence
	AACGTCGACGAGTGGATCGAC GAGCATCACTACCAG	spacer 137 out of 150 /contig 17	ttc	33/36	type I-B / forward	Halorubrum phage CGphi46 genomic sequence
	AACGTCGACGAGTGGATCGAC GAGCATCACTACCAG	spacer 138 out of 150 /contig 17	ttc	33/36	type I-B / forward	Halorubrum phage CGphi46 genomic sequence

Supplementary Table 6: Perfect match cross-targeting spacers from the environmental isolates and Distance from nearest recombinase gene in the contig. All cross-targeting spacers from each isolate are listed with the following description: the name of the isolate from which the

targeting spacer originated from (column "targeting isolate"), the targeted isolate contig sequence, the spacer sequence, the distance in nt from the closest recombinase gene, additional information regarding the recombinase location and type. Identical spacers detected in different isolates are marked in blue.

Targeting isolate	Targeted isolate	Spacer sequence	Distance (in nt) from nearest recombinase gene in the contig	Additional information
7R	47R (contig 5)	AAATCTCGCCGACAGCACCGCTCACATCCCGGCTG	17688	product Tyrosine recombinase XerC;159754
	47R (contig 5)	TGCTGACTCCGGAGTGAGTGTGCACGCCCAGCCATC	153959	
	47R (contig 16)	CTTCGAGGTTCAGCGAGTCCGACCCGCCAGCCGATA	-	Short contig (1258) no recombinase found
	47R (contig 12)	ATCGAGGACGTGCTGGACCGCCACGGCGACTGGTG	-	
	47R (contig 10)	TCACCCACCCAGCAAGTGCGGCAACGGCAACGGCGA	42235	Tyrosine recombinase XerC / xerC_5; 42726
19N	47N (contig 1)	ATGCCGACGACGAGGAGCGCGGTCGTCCAGCCGACC	87839	site-specific tyrosine recombinase XerC; 1137431
24N	48N (contig 6)	TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG	15602	site-specific tyrosine recombinase XerC; 179672
	48N (contig 6)	AAAGACGTACTCTCGCTCGGCGGTCCAGTTCGGCGAC	16827	"
	48N (contig 6)	ACTGAACGAGTCGCCGTTGACTTCCTGAGCGACCTGCC	14449	"
	47N (contig 5)	CTTGATGCGGTCGTTCATCCGGTTGAGACACCGCAGA	58785	site-specific tyrosine recombinase XerC; 120694
47N	48N (contig 6)	ATGGTCAACCTCCTCTGCAACGGGCTCATGACCGC	51505	product Tyrosine recombinase XerC / xerC_2; 135081
	48N (contig 6)	TCACTGCTCATCACCCAACTCAAGAGCCATGACTAACG	15602	"
	48N (contig 6)	GTTGATGACGACAGCGCAGGACTTGGACTCGACTTCAG	23347	"
	48N (contig 1)	TCGCCCGCGGACGCAATCGACGCCGGAATCGGTATGATT	199867	tyrosine recombinase XerC; 112571

Supplementary Table 7: Activity of the nine CRISPR arrays in *H. volcanii* and *H. mediterranei* during inter-species mating. Activity is calculated per array, as the ratio the number of reads with newly acquired spacers (unique or total) to the number of reads mapping to the original array alone or the replicon where the array is located.

Arrays	Number of reads with newly acquired spacers	Number of reads without new spacers	Reads ratio: new spacers / no new spacers	Number of unique newly acquired spacers	Unique ratio: new spacers / no new spacers	Array's replicon representation	Unique normalized by array's replicon representation
A	10026	314959	3.18E-02	3716	1.18E-02	2.90E-01	1.28E+04
B	3397	399004	8.51E-03	2134	5.35E-03	2.90E-01	7.36E+03
C	80355	257628	3.12E-01	52681	2.04E-01	3.97E-01	1.33E+05
D	24292	317609	7.65E-02	14301	4.50E-02	3.97E-01	3.61E+04
I	971	587971	1.65E-03	212	3.61E-04	2.90E-01	7.31E+02
F	2275	428646	5.31E-03	1044	2.44E-03	7.67E-02	1.36E+04
G	538	379338	1.42E-03	268	7.06E-04	7.67E-02	3.50E+03
H	728	368370	1.98E-03	606	1.65E-03	2.90E-01	2.09E+03
E	1	379485	2.64E-06	1	2.64E-06	2.33E-02	4.29E+01

Supplementary Table 8: Comparison of spacer acquisition in *H. volcanii* during inter-species mating, intra-species mating, nutritional competence and at mating conditions without selecting for mating products. The ratio of unique (a) and total (b) newly acquired spacers (targeting **self replicons** only) to the total number of reads that mapped to the original array (i.e. no acquisition) is shown for *H. volcanii's* most active arrays.

a. Unique spacer ratios (new acquisitions/total reads of the original array)

***H. volcanii* arrays**	**Between species mating**	**Within species mating**	**Nutritional competence**	**Mating conditions without selecting for mating products**
C	0.54%	0.0024%	0.05%	0.0007%
D	0.16%	0.046%	0.09%	0.0011%

b. Total spacer-mapped reads ratios (new acquisitions/total reads of the original array)

H. volcanii arrays	**Between species mating**	**Within species mating**	**Nutritional competence**	**Mating conditions without selecting for mating products**
C	6.6%	0.003%	0.064%	0.0007%
D	4.1%	0.14%	0.16%	0.002%

Supplementary Table 9: PAM signature per species across all replicons. For both species, preferred PAM signature are presented both when based on unique spacers and when based on total spacer counts. *H. mediterranei*'s preferred PAM signature (TTC) is marked in red and *H. volcanii*'s preferred PAM (TAC) is marked in blue. *H. mediterranei* replicons are highlighted in orange and *H. volcanii* replicons are in blue.

H. mediterranei

Target	*H. mediterranei* chromosome	pHM100	pHM300	pHM500	*H. volcanii* chromosome	pHV1	pHV3	pHV4
Preferred PAM (unique spacer counts)	38.3% (TTC)	4% (TTC)	32.3% (TTC)	24.3% (TTC)	23.9% (TAC)	39% (TAC)	34.6% (TAC)	61.4% (TAC)
Preferred PAM (total spacer counts)	36.6% (TTC)	11.4% (CGA)	25.9% (TTC)	16.5% (TTC)	21.1% (TAC)	36.7% (TAC)	33.4% (TAC)	60.8% (TAC)

H. volcanii

Target	*H. mediterranei* chromosome	pHM100	pHM300	pHM500	*H. volcanii* chromosome	pHV1	pHV3	pHV4
Preferred PAM (unique spacer counts)	39.6% (TAC)	44.5% (TAC)	68.4% (TAC)	73.5% (TAC)	29% (TAC)	32.3% (TAC)	28% (TAC)	35.4% (TAC)
Preferred PAM (total spacer counts)	45.6% (TAC)	47% (TAC)	69.8% (TAC)	62.1% (TAC)	31.1% (TAC)	40.9% (TAC)	30.9% (TAC)	55.2% (TAC)

Supplementary Table 10: The TAC motif is an effective PAM for *H. volcanii*. Column "pTA409": cells were transformed with the vector and the colonies obtained were counted, Column "pTA409-TAC-P1.1": cells were transformed with the invader plasmid containing the PAM TAC and the colonies obtained were counted. Column "reduction by factor": the number of colonies from the transformation with the invader plasmid (pTA409-TAC-P1.1) and the vector (pTA409) were divided resulting in the factor by which the transformation rate is reduced.

pTA409	pTA409-TAC-P1.1	Reduction by factor
524	1	2.00E-03
1.4	8	6.00E-03
187	1	5.00E-03
1.148	3	3.00E-03
Mean (Standard error); $4.00E^{-3}$ ($9.13E^{-4}$)		

Supplementary Table 11: Strains and plasmids used in this work. For each *Haloferax* strain or plasmid used in this work a specific description is provided specifying the genotype and relevant characteristics.

***Haloferax* strains**	**Description**	**Source/reference**
WR646	*H. mediterranei ΔpyrE2 ΔtrpA*	Lab strain
H729	*H. volcanii ΔhdrB*	[5]
WR510	*H. mediterranei ΔpyrE2*	Lab strain
WR536	*H. volcanii ΔpyrE2 ΔtrpA*	[6]
UG453	*H. mediterranei ΔtrpA*	This work
AN245	*H. mediterranei "targeted strain" ΔpyrE2 ΔtrpA, based on WR646, first spacer from H. volcanii CRISPR loci (P1) with the corresponding PAM sequence is inserted at the ΔtrpA loci*	This work
IT289	*WR646 +first spacer from H. volcanii CRISPR loci (P1 on pHV4 natural plasmid)+PAM inserted at the ΔpyrE2 loci*	This work
IT291	*WR646 +first spacer from H. volcanii CRISPR loci (P1 on pHV4 natural plasmid)+PAM inserted at the ΔpyrE2 and also ΔtrpA loci*	This work
H119	*H. volcanii ΔpyrE2*, Δ*trpA*, Δ*leuB*	[7]
Plasmid	**Description**	**Source**
pTA409	*shuttle vector with pyrE2* marker and pHV1 replication origin	[8]

pTA409-PAM28-P1.1	*spacer P1.1. downstream of PAM28 (TAC)*	9

Supplementary Table 12: DNA oligonucleotides used in this work. DNA primers are listed by name, and their use, sequence, and orientation are provided.

Primer name	Description	primer orientation	Sequence
IS270	Array D , 225 bp	Forward	GGGTCGACGGAAACGTTGAT
IS271		Reverse	AATTGGACCCCGGCTTCG
IS270	Array C , 225 bp	Forward	GGGTCGACGGAAACGTTGAT
IS272		Reverse	TGTGATTCGATACGCGACAC
IS273	Array E , 254 bp	Forward	TGGAACCAATGAACCGGTCG
IS274		Reverse	GGGTCGACGGAAACACTCTT
IS275	Array G , 255 bp	Forward	TAACACTCTCGGTTGATGGGG
IS276		Reverse	AGGTCGACGGAAACTGTTGA
IS278	Array F , 258 bp	Forward	GGAGGTCGACGGAAACACTT

IS278		Reverse	TGAGCGACCCAATCGTCTTC
IS279	Array H , 255 bp	Forward	GGGGTCGACGGAAACTGTTG
IS280		Reverse	GGGGTGGGTGGTGATGGAC
IS281	Array B , 254 bp	Forward	ACGGAAACTGTTGAGTGGGA
IS282		Reverse	CGGATGCGACTGTCTGACG
IS283	Array A , 256 bp	Forward	CGTGGGCGACCTCGTA
IS284		Reverse	ACGGAAACACTTGAGTGCGA
IS386	Array I , 353 bp	Forward	TCAAGGACTCGCTGTTGTCG
IS387		Reverse	CTGTAACCGTCGGGAAGGC
IS15	*H. mediterranei pyrE2*	up- Forward	AAAATCTAGACTGCGTTCCCCGGTCTC GACC
IS16	*H. mediterranei pyrE2*+proto spacer	down-Reverse	CAAAGTGTTCCGGGAGGTCGCCGGTC GAGATGCCTGCGAAGGTTGGTTCGGG GCGAAGTAGG
IS17	*H. mediterranei pyrE2*+ proto spacer	down-Forward	TTCGCAGGCATCTCGACCGGCGACCTC CCGGAACACTTTGCTCTTCGAAACCGA CCCGCGG
IS18	*H. mediterranei pyrE2*	down-Reverse	AAAAAAGCTTCAGTTCGATGTCGTGTT CGGCC
IS19	Verifying proto-spacer entered in *ΔpyrE2*	Forward	TCTTGCACGTCGTCACCAA
IS20		Reverse	AGAGCAAAGTGTTCCGGGAG
IS25	Verifying proto-spacer in *ΔtrpA*	Forward	ATTCGCAGGCATCTCGACC
IS26		Reverse	GAGATTTGCGAGTTGCGTCA

AP357	used in building AN245	Forward	AAAAGATCTGTAAAACGACGGCCAGT
AP358		Reverse	AAAAGATCTAACAGCTATGACCATGATTACG
AP381	Verifying proto-spacer entered in *ΔtrpA*	Forward	CAAAGTGTTCCGGGAGGTCGC
IS415	*H. mediterranei* Island 1 flanking regions	Forward	CACGACACTCATCGGGTCAA
IS416		Reverse	AACCGGGGTACGGATGAATG
IS417	*H. mediterranei* Island 1 circular	Forward	CGGGTTCGAGGAGACAGATG
IS418		Reverse	CGTCTCGTCGTCTTCGGTG

Supplementary References

1. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8,** 172 (2007).
2. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497–3500 (2003).
3. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26,** 335–340 (2010).
4. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17,** 356 (2016).
5. Naor, A., Lapierre, P., Mevarech, M., Papke, R. T. & Gophna, U. Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr. Biol.* **22,** 1444–1448 (2012).
6. Abu-Qarn, M., Eichler, J. & Sharon, N. Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea. *Curr. Opin. Struct. Biol.* **18,** 544–550 (2008).
7. Allers, T., Ngo, H. P., Mevarech, M. & Lloyd, R. G. Development of Additional Selectable Markers for the Halophilic Archaeon Haloferax volcanii Based on the leuB and trpA Genes. *Appl. Environ. Microbiol.* **70,** 943–953 (2004).
8. Delmas, S., Shunburne, L., Ngo, H.-P., Allers, T. & Myllykallio, H. Mre11-Rad50 Promotes Rapid Repair of DNA Damage in the Polyploid Archaeon Haloferax volcanii by Restraining Homologous Recombination. *PLoS Genet.* **5,** e1000552 (2009).
9. Fischer, S. *et al.* An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.* **287,** 33351–33365 (2012).