Prophages are associated with extensive CRISPR-Cas auto-immunity

Franklin L. Nobrega[1,2,#], Hielke Walinga[1,2], Bas E. Dutilh[3,*], and Stan J.J. Brouns[1,2,*]

[1] Department of Bionanoscience, Delft University of Technology, Delft, Netherlands

[2] Kavli Institute of Nanoscience, Delft, Netherlands

[3] Theoretical Biology and Bioinformatics, Science4Life, Utrecht University, Utrecht, Netherlands

* To whom correspondence should be addressed. Email: stanbrouns@gmail.com, bedutilh@gmail.com

# Current address: School of Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK

ABSTRACT

CRISPR-Cas systems require discriminating self from non-self DNA during adaptation and interference. Yet, multiple cases have been reported of bacteria containing self-targeting spacers (STS), i.e. CRISPR spacers targeting protospacers on the same genome. STS has been suggested to reflect potential auto-immunity as an unwanted side effect of CRISPR-Cas defense, or a regulatory mechanism for gene expression. Here we investigated the incidence, distribution, and evasion of STS in over 100,000 bacterial genomes. We found STS in all CRISPR-Cas types and in one fifth of all CRISPR-carrying bacteria. Notably, up to 40% of I-B and I-F CRISPR-Cas systems contained STS. We observed that STS-containing genomes almost always carry a prophage and that STS map to prophage regions in more than half of the cases. Despite carrying STS, genetic deterioration of CRISPR-Cas systems appears to be rare, suggesting a level of escape from the potentially deleterious effects of STS by other mechanisms such as anti-CRISPR proteins and CRISPR target mutations. We propose a scenario where it is common to acquire an STS against a prophage, and this may trigger more extensive STS buildup by primed spacer acquisition in type I systems, without detrimental autoimmunity effects. The mechanisms of auto-immunity evasion create tolerance to STS-targeted prophages, and contribute both to viral dissemination and bacterial diversification.

Keywords: CRISPR-Cas; Auto-immunity; Self-targeting; Anti-CRISPR protein; Escape; Bacteriophage; Prophage; Transposon

INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated proteins (Cas) are defense systems, which provide bacteria and archaea with an adaptive and heritable immunity against invading genetic elements such as bacteriophages or plasmids (1-3). Immunity is conferred by small sequences, known as spacers, which are taken up from the invaders' genome and integrated into the CRISPR locus (2). At the CRISPR locus, spacers function as the system's memory, and are used in the form of guide RNA to specifically recognize and degrade foreign DNA or RNA (3-5). While known to be highly specific for their target, CRISPR-Cas systems do pose a risk for auto-immunity if spacers from the host chromosome are mistakenly acquired (6). These self-targeting spacers (STS) have been reported in numerous species, and their most likely consequence is cell death by directing cleavage and subsequent degradation of the host genome (7,8). Escape from the lethal outcome of auto-immunity occurs for cells selected for mutations on the target sequence (9,10) and/or for inactivation of CRISPR-Cas functionality via, for example, mutation or deletion of the Cas genes, spacers, repeats, or protospacer adjacent motifs (PAM). The action of anti-CRISPR (Acr) proteins encoded by prophages may also prevent auto-immunity (11). In fact, the presence of STS in a genome has been suggested (11,12) and recently successfully employed (13) as a strategy to discover new Acrs.

Auto-immunity has been mostly regarded as a collateral effect of CRISPR-Cas systems, but it has also been suggested to play a role in the evolution of bacterial genomes on a population level by influencing genome remodeling (9). Although reported only on isolated examples, CRISPR-Cas systems have been speculated to act like a regulatory mechanism (14-17). Auto-immunity has also been proposed to be triggered by foreign DNA with similarity to the bacterial chromosome (18).

Here we take a closer look at STS in the many types and subtypes of CRISPR-Cas systems to identify the incidence, distribution and mechanism of evasion of potential CRISPR-Cas auto-immunity in bacteria. We demonstrate that STS are frequently observed in bacterial genomes, and that bacteria have evolved mechanisms to evade death by auto-immunity while preserving their CRISPR-Cas systems. We propose that the integration of phages in the bacterial chromosome provides evolutionary advantages to the bacteria (e.g. acquisition of virulence traits) but is also the primary trigger of STS acquisition in CRISPR arrays. We further suggest that mechanisms of evasion from auto-immunity create tolerance to the integrated invaders, benefiting both bacteria and phage populations by allowing the acquisition of novel genetic information by the bacteria, and by promoting phage (passive) dissemination in the bacterial population.

Material and methods

Detection of CRISPR arrays

The complete genome collection of the PATRIC database (19) (a total of 110,334 genomes) was used in our analysis. CRISPR arrays were predicted for each genome using CRISPRDetect 2.2.1 (20) with a quality score cut-off of 3.

Detection of self-targeting spacers

All spacers were blasted (blastn-short option, DUST disabled, e-value cut-off of 1, gap open, and gap extend penalty of 10) against the source genome. The blastn results were filtered for a minimum identity higher than 90% with the target. Any hit on the genome was considered a self-target, except for those within all of the predicted CRISPR arrays, including arrays identified with a CRISPRDetect quality score below 3. Hits closer than 500 bp from each end of the predicted arrays were also ignored to avoid considering spacers from the array that were possibly not identified by CRISPRDetect. Spacers with flanking repeats of identity score lower than 75% to each other were discarded as they may have been erroneously identified as spacers. Of these, only spacers smaller than 70 bp and a repeat size between 24 and 50 bp were retained in the dataset. Finally, STS from CRISPR arrays of two or fewer spacers were excluded, except when the associated repeat belonged to a known CRISPR repeat family, as identified by CRISPRDetect. Duplicates were removed by search of similar genomes, contigs and arrays.

Classification of CRISPR-Cas systems

The CRISPR-Cas systems of STS-containing genomes were classified using MacsyFinder (21) in combination with Prodigal (22), and the CRISPR-type definitions and Hidden Markov Models (HMM) profiles of CRISPRCasFinder (23). The classification of the repeat family of the CRISPR array was obtained using CRISPRDetect. Genomes carrying two or more CRISPR-Cas types were labeled as ‘mixed’, and those having CRISPR-Cas arrays but no *cas* genes were labeled as ‘no Cas’. Systems which could not be assigned a

CRISPR sub-type and which were missing at least one *cas* gene (but contained no less than one *cas* gene) were classified as 'incomplete'. The final classification of each genome can be found in Supplementary Table S1.

Analysis of the genomic target

The orientation of the arrays was determined by CRISPRDetect using the default parameters of CRISPRDirection. After this, the STS sequence was used for a gapless blastn at the target and to retrieve the PAM downstream or upstream of the STS based on the CRISPRDetect classification (see Supplementary Table S1). The targets were then analyzed for the correct PAM sequence by comparison with the expected PAM for the different CRISPR-Cas types as previously described (11,24,25). The consensus PAM sequences used in this analysis are shown in Supplementary Table S2. Genes of STS-containing genomes were predicted using Prodigal and annotated using Interproscan (26) and Pfam (27) domain prediction. Prophage regions in the genomes were detected using VirSorter (28), and used to identify STS targeting these regions. Transposons were also detected in the genomes using Interproscan (26) (Supplementary Table S3). Targets of the STS with e-value $<10^{-5}$ were grouped by function to identify possibly enriched hits separately for prophage and endogenous regions. Only those hits associated with predicted correct PAMs were subjected to this analysis.

Distance between self-targeting spacer and prophages

Contigs predicted to contain prophages were extracted and used to create a hit density map based on STS distance to prophage(s).

Identification of anti-CRISPR proteins

The amino acid sequences of known Acrs (29) were used for homology search in the STS-containing genomes using BLASTp with an e-value limit of 10^{-5}.

Statistical analysis

A binomial test was performed on CRISPR arrays of different sizes to test the hypothesis that STS at the leader side of the CRISPR array are more common. Only STS from CRISPR arrays smaller than 50 spacers were considered because larger arrays are too scarce to result in a reliable statistical analysis. A chi-squared test was used to determine statistical significance between percentages of populations. Statistical significance was considered for *P* < 0.05.

Software

GNU parallel was used to parallelize tool runs and for parsing of output files (30). Biopython package (31) functions were used for specific analysis, such as GFF parser for prodigal files, pairwise2 for removing false positives based on repeat identity, and nt_search for matching of the PAM. All data collected was managed using Python package Pandas (32). Python packages SciPy (33), Matplotlib (34) and Seaborn (35) were used for statistical analysis and visualization.

Results

Self-targeting spacers (STS) are often found in CRISPR-encoding bacteria

We scanned 43,526 CRISPR-encoding genomes for spacers with >90% sequence identity to the endogenous genomic sequence that is not part of a CRISPR array. We decided upon this definition of STS as a 10% mismatch between spacer and target can still trigger a functional CRISPR response (direct interference and/or priming in type I) in many CRISPR-Cas types (36-41). For clarity, we note that our definition of STS may exclude or include certain sequences as a result. For example, STS protospacers that suffered extensive mutations may be excluded, while spacers that target non-genomic regions of high similarity to a genomic region may be included. We found that 23,626 out of 1,481,476 spacers (1.6%) are self-targeting based on this cutoff. Approximately half of those (12,121, 0.8%) had 100% sequence identity to the genome from which the spacers were derived (frequency of STS with mismatches can be seen in Supplementary Table S4), a percentage higher than previously reported (0.4% with 100% identity) (14). Similar to previous observations with smaller datasets (14), about one fifth (19%, 8,466) of CRISPR-encoding genomes have at least one STS in one of their CRISPR arrays.

We further looked into how frequent STS were in different types of CRISPR-Cas systems (Figure 1A). STS were detected in genomes containing CRISPR-Cas systems of almost all subtypes, and were more prevalent (>40%) in CRISPR-Cas types I-B and I-F. Curiously, genomes containing STS are almost absent in type III-A, but present between 10 and 20%

in type III-B, C and D systems. Moreover, length of the STS agreed with reported preferred spacer length for different CRISPR-Cas subtypes (Supplementary Figure S1) (42-44).

It has been suggested that following the integration of an STS, the CRISPR-Cas system must become inactivated in order to survive, and that this phenomenon could explain the abundance of highly degraded CRISPR systems that contain *cas* pseudogenes (14). Recent experimental evolution studies have shown that large genomic deletions encompassing the entire CRISPR-Cas locus can occur as a consequence of auto-immunity to prophages (45). We observed that 12% (979 of 8,466) of the STS-containing genomes contain incomplete CRISPR systems or no *cas* genes, while 88% (7,490 of 8,466) seem to carry intact CRISPR-Cas systems ($P < 0.0001$, chi-squared t-test, Figure 1A). This suggests that CRISPR-Cas deletion can occur as a mechanism to survive STS, but self-targeting can also be overcome through other mechanisms. To note that our homology-based analysis cannot account for small inactivating mutations in *cas* genes that could also render a CRISPR-Cas system non-functional, but we expect that the effect of such recent pseudogenization is minor as inactive pseudogenes tend to be rapidly lost from the genome (46,47). Moreover, we found that most STS locate in the leader proximal positions of the array (Figure 1B, Supplementary Figure S2), with several STS also found in middle and leader distal positions (Figure 1B). To account for potential bias introduced in this analysis by smaller arrays, we generated the same plot for arrays of 10 or less spacers (Supplementary Figure S3). The same trend is apparent, confirming that STS preferably locate near the leader but are also present in later positions in the array. This suggests that the CRISPR system (or at least memory acquisition) remains active after integration of an STS into the CRISPR array and the cell remains viable. Correct CRISPR array orientation prediction remains challenging in some cases (48), and there may be some arrays in our database whose orientation was predicted incorrectly by the CRISPRDirection tool. This may lead to noise in the positionality of STS. Still, we are confident on our overall observations as CRISPRDirection is backed up by experimental evidence for most CRISPR types, including type I-U (49).

In summary, STS are common among bacteria harboring all types of CRISPR-Cas systems, but especially types I-B and I-F. Importantly, STS-containing bacteria seem to preserve CRISPR-Cas, perhaps by employing alternative mechanisms to avoid the lethal effects of auto-immunity.

STS are enriched in prophage-containing genomes

To understand if targeting of endogenous regions by STS could have a regulatory role in gene expression, we looked at the position of STS hits in the genome and determined if these were in coding or non-coding regions. In general, no preference for targeting non-coding regions was observed, with coding regions being predominant in most types of CRISPR-Cas systems (*P* < 0.05, chi-squared test, Supplementary Table S5), with the exception of STS of types I-D, III-A, III-B, V-B and VI-A CRISPR-Cas systems for which intergenic and coding regions are equally targeted (*P* > 0.05, Figure 2A, Supplementary Table S5). This suggests that there is no apparent link between CRISPR-Cas auto-immunity and regulating promoter activity for gene expression. Still, no absolute conclusions can be drawn about a potential regulatory role of STS since direct targeting of genes (coding regions) leads to programmed regulation of gene expression (50-53). Also, in most cases we could not detect a preference for targets on the sense or antisense DNA strands (*P* > 0.05, Figure 2A, Supplementary Table S5).

Bacteriophages are common targets of CRISPR-Cas systems and exist abundantly in nature. Because some bacteriophages can integrate into the bacterial chromosome, we next investigated if the presence of prophages in a genome would associate with the presence of STS. We identified prophage regions in the STS-containing genomes and observed that, on average, 52.4% of the STS-containing genomes have STS with protospacers in prophage regions, with type I-F CRISPR-Cas systems showing up to 70% genomes with prophage hits (Figure 2B). Interestingly, we also observed that 96.9% (8,203 out of 8,466) of the STS-containing genomes have at least one integrated prophage, while only 28.5% (9,992 out of 35,060) of the STS-free genomes contain prophages (*P* < 0.0001, chi-squared test). It therefore appears that STS is linked to carrying prophages.

We further questioned if STS were also enriched in bacteria containing other mobile genetic elements able to integrate into the bacterial genome. To do so, we looked at the prevalence of transposons in STS-containing and STS-free genomes of bacteria with CRISPR arrays. We observed a moderately higher prevalence of transposons in STS-containing genomes (12.1% vs 7.7%, or 10.9% vs 5.0% when discarding incomplete and no Cas genomes, *P* = 0.004 and *P* = 0.001, respectively, chi-squared test) (Figure 2C).

We next wondered if collateral targeting of prophage regions would lead to STS of endogenous genomic regions flanking the prophage. To test this we mapped the distance of STS in the genome to the nearest prophage region. For this we considered only STS targeting regions of complete genomes and contigs which contained a prophage. 59.5% of these STS target a prophage region, while the remainder mostly target the nearby endogenous genome (Figure 2D). Distances to prophage were also normalized by contig length to discard possible variations due to differences in contig size, which shows a similar pattern of STS hitting regions close to the prophage (Supplementary Figure S4). This suggests that targeting of endogenous regions is indeed related to proximity to a prophage region. As the definition of prophage boundaries may be associated with a certain level of inaccuracy, nearby STS protospacers may also be part of the prophage itself. Because genomic regions flanking prophages are often excised together with the prophage, it is also possible that such regions are subjected to spacer acquisition when the prophage enters its lytic cycle. Finally, prophages tend to repeatedly integrate in the same regions of bacterial genomes, so it is possible that proximal prophage regions are enriched in degenerated prophages as well. All these processes could contribute to the enrichment of STS in prophages and their proximal genomic regions, as shown by our results.

In summary, 63% of STS are linked to prophages or the nearby endogenous genome (<50 kb, see Figure 2D). Thus, our data suggest that the occurrence of STS is strongly linked to the presence of prophages in the bacterial chromosome.

Interference-functional STS with consensus PAM are frequent in type I CRISPR-Cas systems

To explain how STS are tolerated we first looked at the targeting requirements of CRISPR-Cas systems. In many CRISPR-Cas systems, the correct identification of the target is dependent on a small 2-6 base pair motif immediately adjacent to the target DNA sequence, known as the PAM (54). The PAM is essential for binding to and cleavage of the target DNA by the Cas nucleases, and mutations in this sequence can abrogate targeting (55). To understand how often STS protospacers have a consensus PAM, and can therefore be efficiently targeted, we compared the PAM sequence of the STS protospacer with the expected PAM sequence for the different CRISPR-Cas types previously described (Supplementary Table S2) (24,25,56). We observed that 22.4% of all STS (4,140 of 18,483 STS with 90% sequence identity) and 23.9% of STS with 100 % identity (2,294 of 9,605) have a consensus PAM (Figure 4A and Supplementary Table S6), suggesting these to be functional for direct interference. Type I CRISPR-Cas systems, especially types I-B (29.5%), I-C (44.7%) and I-E (37.0%) have more STS with a consensus PAM (average 27.5%) than type II (average 0.1%) or type V (average 12.8%) (Figure 4A). This may suggest that bacteria encoding type II and type V systems avoid the lethal effects of auto-immunity by having non-functional STS, while bacteria encoding type I systems may employ other evasion mechanisms to withstand the lethal auto-immunity effects of interference-functional STS.

Several factors should be considered when analyzing the role of PAM sequences in tolerance mechanisms to STS. First, the full diversity of functional PAM sequences in nature currently remains unknown, as does their distribution across taxa. Second, PAM sequences can vary widely even within a CRISPR subtype (e.g. in different species) (54,57-59). Third, different CRISPR class I (type I, III and IV) systems may use different PAM sequences for spacer acquisition and for targeting (60). Our analysis has revealed a range of candidate bacteria that can contain mechanisms allowing them to remain viable while carrying interference-functional STS with known consensus PAM sequences. It will be interesting to see these mechanisms further unraveled in future studies.

Acrs are more prevalent in bacteria carrying STS

To understand how bacteria are able to survive STS while keeping their *cas* genes intact, we assessed the presence of Acrs encoded by prophages. By inhibiting the activity of the CRISPR-Cas system using a variety of mechanisms (reviewed in (29)), Acrs can prevent the lethal effects of STS auto-immunity. In fact, STS have been used to identify new Acr proteins (13,61). We mapped Acrs in the STS-containing genomes using homology searches with all currently known Acrs (29). Acrs were found at low frequency (10.9% average, Figure 4B) but still at levels significantly higher than those found in STS-free, CRISPR-containing genomes (0.3% average, $P < 0.0001$, chi-squared test). The levels of Acrs here reported are a lower bound, as unidentified Acrs may be present in these genomes and these proteins may thus have a higher influence in escaping auto-immunity. Even so, we found many Acr homologs in STS-containing bacteria carrying single type I-B, IV or VI-A CRISPR-Cas systems, for which no Acrs have yet been described (Figure 4B and Supplementary Table S7). Putative Acrs for type I-B and type IV CRISPR-Cas systems were recently identified by using a bioinformatics pipeline (61), but to our knowledge none has yet been suggested for type VI-A.

Among the newly found Acrs, homologs of AcrIF2-7, AcrIF11-13 and AcrIIA1-4 were the most common in STS-containing genomes (Figure 4C). Interestingly, homologs of AcrIF1-14, AcrIE1-5, and AcrIIA1-4 were found in genomes of diverse CRISPR-Cas subtypes, while homologs of AcrVA1-5 and AcrIIC2-5 appear only in genomes containing the corresponding CRISPR-Cas subtype. Particularly, homologs of AcrIF1-14 and AcrIE1-5 were found in type I and type IV CRISPR-Cas types, while homologs of AcrIIA1-4 were detected in type I, II and VI-A CRISPR-Cas systems. Acr homologs of families that do not correspond to the CRISPR-Cas system found in the bacteria were also recently reported (61). It is possible that some Acr homologs have activity against multiple types of CRISPR-Cas systems, which may occur if the mechanism of inhibition of the Acr is compatible with the multiple types. The ability of Acrs to inhibit different types of CRISPR-Cas systems has already been revealed for some Acrs (62,63), although the specific mechanisms of inhibition have not yet been described.

Anti-CRISPR associated (aca) genes were also found, especially in types I-E and I-F CRISPR-Cas systems, and with higher prevalence of aca1 (488) and aca4 (220) genes (see Supplementary Figure S6 and Supplementary Table S7).

In conclusion, among genomes with a CRISPR system, Acrs are more prevalent in genomes containing STS than in genomes without STS, and it therefore is likely that Acrs play a major role in auto-immunity evasion.

Amplified self-targeting in prophages regions

In our analysis, we found 1,224 genomes with a number of STS higher than the average (2.5 ± 2.9 STS, Supplementary Figure S5). We decided to take a closer look at two extreme cases and investigate how STS with 100% identity were distributed in the bacterial chromosome (Figure 3). The genome of *Blautia producta* strain ATCC 27340 contains a type I-C CRISPR-Cas system and 11 prophage regions in the chromosome (Figure 3A). This strain contains a stunning 162 STS mostly hitting prophage regions. The genome of *Megasphaera elsdenii* strain DSM 20460 contains three distinct CRISPR-Cas systems (types I-C, I-F and III-A), two large prophage regions (Figure 3B) and a total of 85 STS in its I-C CRISPR arrays. In *B. producta* and *M. elsdenii*, the wealth of STS hit mostly in and around prophage regions, with some prophages remaining untargeted. After manual confirmation of the consensus repeat and array orientation of the STS, we observed that the oldest STS (located further from the leader in the CRISPR array) are those with protospacer in the prophage regions (Figure 3A and 3B), suggesting these were the initial hits and that additional spacers could have been acquired from locations in the prophage vicinity by primed CRISPR adaptation. Interestingly, as priming is enhanced by CRISPR interference (18,64,65), it is striking that no apparent DNA damage was incurred. For *M. elsdenii* we found that all STS protospacers are on the same strand with an orientation bias characteristic of primed adaptation (18). Primed adaptation would result in the acquisition of many spacers, explaining the high number of STS found in these genomes. It is interesting that STS in *M. elsdenii* were integrated in only two out of six CRISPR arrays, both close to the I-C *cas* genes (Figure 3B). It is also curious to note that no homologs of any known Acr (29) could be found in either genome using BLASTp homology searches with an e-value cutoff of 10^{-5}.

Overall, these examples of extensive, tolerated self-targeting suggest that prophage integration was followed by primed adaptation, leading to the amplification of STS against the prophage and flanking genomic regions.

DISCUSSION

Self-targeting CRISPR spacers (STS) in bacteria are not a rare phenomenon, as one fifth of bacteria with CRISPR systems carries STS. Interestingly, some types of CRISPR-Cas systems (i.e. types I-B and I-F) seem to be more prone to incorporation of STS into CRISPR arrays. As STS may lead to auto-immunity, here we questioned which mechanisms could drive STS acquisition and whether bacteria encode mechanisms to protect themselves. We observed a striking prevalence of prophages in STS-containing genomes when compared to STS-free genomes, suggesting that prophages could be the trigger of STS acquisition. Only about half of the STS targeted protospacers are located within the prophage regions, with the other half targeting the endogenous genome. Interestingly, STS hits in the endogenous genome are enriched in the proximity of prophages, showing a pattern consistent with primed adaptation from an initial protospacer present on the prophage. Also, in cases where bacteria carried multiple STS, the STS located the furthest from the leader sequence targeted the prophage, while subsequent STS targeted both prophage and endogenous regions. These results are consistent with a model where primed adaptation amplifies STS by acquisition of new spacers from both prophage and prophage-adjacent regions.

STS can lead to lethal auto-immunity, but we still found many STS-containing bacteria in the genome database, as well as many STS functional for direct interference (associated with a consensus PAM) capable of efficient targeting, especially in type I CRISPR-Cas systems. This suggests bacteria employ other mechanisms of auto-immunity evasion to survive. Interestingly, degradation of the CRISPR-Cas system itself does not seem to be the dominant evasion mechanism employed by bacteria to survive potential auto-immunity caused by STS, as we found at least 4 times more genomes with intact rather than degraded CRISPR-Cas systems. Genomes carrying type II and V CRISPR systems commonly have non-consensus PAM sequences of the STS protospacer which may help avoid auto-immunity. Whether this occurs by incorrect acquisition of the spacer (66,67), or mutation of the PAM when it is already integrated, is unknown. Although found at low frequency, Acrs were also present significantly (36-fold) more often in STS-containing genomes than STS-free genomes.

Based on our overall observations, we here suggest two scenarios for the appearance of STS in bacterial genomes. In the first scenario, bacteria may acquire a first spacer against a temperate phage, but despite this, the phage may still be able to integrate into the genome. In the second scenario, a prophage may already be integrated into the genome and the ‘accidental’ acquisition of an STS by the host may start targeting the prophage. Following this first STS, incomplete targeting may lead to further STS expansion by primed spacer acquisition, in type I and II systems (68,69), which will result in the incorporation of multiple new spacers targeting both the prophage and adjacent locations in the bacterial genome. This continuation of extensive priming, which is thought to require a level of CRISPR targeting, is without apparent genome damage or lethality. The process of acquiring STS creates an apparent standoff between CRISPR-Cas and targeted prophages that involves mechanisms of auto-immunity avoidance and anti-phage defense. As shown here, these interactions may involve Acrs that may contribute to creating tolerance to STS in general, and to STS-targeted prophages in particular. Thus, it is possible that the CRISPR system may be preventing prophage induction (70), and perhaps induce prophage clearance or genome deletions (71,72). When the protospacer region of the prophage in the bacterial genome is deleted, this may lead to interesting eco-evolutionary dynamics, as the presence of the former STS on the bacterial genome may now prevent reinfection of the immunized strain by the same or related phages. Similarly, if the CRISPR system prevents induction of the prophage by targeting it upon excision from the genome, the induction of the lytic cycle could be inhibited and the shift from lysogeny to a lytic state could be detected and acted upon. The balance between these processes remains subject to further experimentation and modelling.

It has been suggested that CRISPR-Cas systems could have some tolerance to mobile genetic elements to allow acquisition of potentially beneficial genetic information (73). Tolerance to prophages has been observed, but not to plasmids (73,74). Maintenance of a plasmid bearing beneficial traits in specific environmental contexts has been shown to lead to CRISPR loss (75,76), although probably resulting from the beneficial plasmid helping select for cells without CRISPR-Cas that could randomly appear in the population rather than the plasmid actively causing CRISPR-Cas loss. Tolerance may not be equal to all mobile genetic elements, such as mobile genetic elements that integrate the bacterial chromosome (e.g. prophages and transposons) as a consequence of the presence of Acrs or of selection for different modes of escape from self-targeting. Tolerance to integrated mobile genetic elements derived from auto-immunity escape may breach the barrier imposed by CRISPR-Cas systems and facilitate the diversification and evolution of bacterial genomes and the passive dissemination of phages in bacterial populations.

AVAILABILITY

Data is available in the GitHub repository (https://github.com/hwalinga/self-targeting-spacers-scripts and https://github.com/hwalinga/self-targeting-spacers-notebooks).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

This work was supported by the Netherlands Organisation for Scientific Research (NWO) [grant numbers VENI 016.Veni.181.092 to F.L.N., VIDI 864.14.004 to B.E.D., and VICI VI.C.182.027 to S.J.J.B.]; and European Research Council (ERC) Stg grant [grant number 639707 to S.J.J.B.] and Consolidator grant [grant number 865694 to B.E.D]. Funding for open access charge: NWO [grant number 016.Veni.181.092].

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Jansen, R., Embden, J.D.A.v., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565-1575.
2. Mojica, F.J.M., Díez-Villaseñor, C.s., García-Martínez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*, **60**, 174-182.
3. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709-1712.
4. Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V. and van der Oost, J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960-964.
5. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. and Terns, M.P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, **139**, 945-956.
6. Wimmer, F. and Beisel, C.L. (2020) CRISPR-Cas systems and the paradox of self-targeting spacers. *Front Microbiol*, **10**, 3078-3078.
7. Heussler, G.E. and O'Toole, G.A. (2016) Friendly Fire: Biological functions and consequences of chromosomal targeting by CRISPR-Cas systems. *J Bacteriol*, **198**, 1481-1486.

8. Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V. and Koonin, E.V. (2017) The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *mBio*, **8**, e01397-01317.

9. Selle, K., Klaenhammer, T.R. and Barrangou, R. (2015) CRISPR-based screening of genomic island excision events in bacteria. *Proc Natl Acad Sci U S A*, **112**, 8076-8081.

10. Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., Chen, L., Liang, Y.X. and She, Q. (2016) Harnessing Type I and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Res*, **44**, e34-e34.

11. Rauch, B.J., Silvis, M.R., Hultquist, J.F., Waters, C.S., McGregor, M.J., Krogan, N.J. and Bondy-Denomy, J. (2017) Inhibition of CRISPR-Cas9 with bacteriophage proteins. *Cell*, **168**, 150-158.e110.

12. Bondy-Denomy, J. (2018) Protein inhibitors of CRISPR-Cas9. *ACS Chem Biol*, **13**, 417-423.

13. Watters, K.E., Fellmann, C., Bai, H.B., Ren, S.M. and Doudna, J.A. (2018) Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science*, **362**, 236-239.

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

15. Cady, K.C. and O'Toole, G.A. (2011) Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *J Bacteriol*, **193**, 3433-3445.

16. Li, R., Fang, L., Tan, S., Yu, M., Li, X., He, S., Wei, Y., Li, G., Jiang, J. and Wu, M. (2016) Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Res*, **26**, 1273-1287.

17. Ratner, H.K., Escalera-Maurer, A., Le Rhun, A., Jaggavarapu, S., Wozniak, J.E., Crispell, E.K., Charpentier, E. and Weiss, D.S. (2019) Catalytically active Cas9 mediates transcriptional interference to facilitate bacterial virulence. *Mol Cell*, **75**, 498-510.e495.

18. Staals, R.H.J., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M. and Fineran, P.C. (2016) Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun*, **7**, 12853-12853.

19. Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L. *et al.* (2016) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res*, **45**, D535-D542.

20. Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C. and Brown, C.M. (2016) CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, **17**, 356-356.
21. Abby, S.S., Néron, B., Ménager, H., Touchon, M. and Rocha, E.P.C. (2014) MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE*, **9**, e110726.
22. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
23. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D. and Pourcel, C. (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res*, **46**, W246-W251.
24. Westra, E.R., Swarts, D.C., Staals, R.H.J., Jore, M.M., Brouns, S.J.J. and Oost, J.v.d. (2012) The CRISPRs, They are a-changin': How Prokaryotes Generate Adaptive Immunity. *Annu Rev Genet*, **46**, 311-339.
25. Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G. and Randau, L. (2019) PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA Biol*, **16**, 504-517.
26. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236-1240.
27. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res*, **42**, D222-D230.
28. Roux, S., Enault, F., Hurwitz, B. and Sullivan, M. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
29. Trasanidou, D., Gerós, A.S., Mohanraju, P., Nieuwenweg, A.C., Nobrega, F.L. and Staals, R.H.J. (2019) Keeping crispr in check: diverse mechanisms of phage-encoded anti-crisprs. *FEMS Microbiol Lett*, **366**, fnz098.
30. Tange, O. (2011) GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine*, **February 2011**, 42-47.

31. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422-1423.
32. McKinney, W. (2010) Data structures for statistical computing in Python. *Proc of the 9th Python in Science Conf (SCIPY 2010)*, 51-56.
33. Jones, E., Oliphant, T. and Peterson, P. (2001) *SciPy: Open Source Scientific Tools for Python*.
34. Hunter, J.D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, **9**, 90-95.
35. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D., Augspurger, T., Halchenko, Y., Cole, J. *et al.* (2018) mwaskom/seaborn: v0.9.0 (July 2018). *Zenodo*, http://doi.org/10.5281/zenodo.1313201.
36. Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A.F.T., Staals, R.H.J. and Brouns, S.J.J. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A*, **111**, E1629-E1638.
37. Goldberg, G.W., McMillan, E.A., Varble, A., Modell, J.W., Samai, P., Jiang, W. and Marraffini, L.A. (2018) Incomplete prophage tolerance by type III-A CRISPR-Cas systems reduces the fitness of lysogenic hosts. *Nat Commun*, **9**, 61-61.
38. Manica, A., Zebec, Z., Steinkellner, J. and Schleper, C. (2013) Unexpectedly broad target recognition of the CRISPR-mediated virus defence system in the archaeon Sulfolobus solfataricus. *Nucleic Acids Res*, **41**, 10509-10517.
39. Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J. and Severinov, K. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A*, **108**, 10098-10103.
40. Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A. and Liu, D.R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol*, **31**, 839-843.
41. Zhang, B., Ye, Y., Ye, W., Perčulija, V., Jiang, H., Chen, Y., Li, Y., Chen, J., Lin, J., Wang, S. *et al.* (2019) Two HEPN domains dictate CRISPR RNA maturation and target cleavage in Cas13d. *Nat Commun*, **10**, 2544.

42. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. and Wang, Y. (2015) Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell*, **163**, 840-853.
43. Kuznedelov, K., Mekler, V., Lemak, S., Tokmina-Lukaszewska, M., Datsenko, K.A., Jain, I., Savitskaya, E., Mallon, J., Shmakov, S., Bothner, B. *et al.* (2016) Altered stoichiometry Escherichia coli Cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. *Nucleic Acids Res*, **44**, 10849-10861.
44. Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N. and Doudna, J.A. (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, **527**, 535-538.
45. Rollie, C., Chevallereau, A., Watson, B.N.J., Chyou, T.-y., Fradet, O., McLeod, I., Fineran, P.C., Brown, C.M., Gandon, S. and Westra, E.R. (2020) Targeting of temperate phages drives loss of type I CRISPR–Cas systems. *Nature*, **578**, 149-153.
46. Mira, A., Ochman, H. and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet*, **17**, 589-596.
47. Kuo, C.-H., Moran, N.A. and Ochman, H. (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res*, **19**, 1450-1454.
48. Milicevic, O., Repac, J., Bozic, B., Djordjevic, M. and Djordjevic, M. (2019) A simple criterion for inferring CRISPR array direction. *Front Microbiol*, **10**.
49. Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J J. (2019) Cas4–Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res*, **47**, 5223-5230.
50. Luo, M.L., Mullis, A.S., Leenay, R.T. and Beisel, C.L. (2015) Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic Acids Res*, **43**, 674-681.
51. Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R. and Lundgren, M. (2015) Efficient programmable gene silencing by Cascade. *Nucleic Acids Res*, **43**, 237-246.
52. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P. and Lim, W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173-1183.
53. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L.A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res*, **41**, 7429-7437.

54. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J. and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733-740.

55. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C. and Brouns, S.J.J. (2017) CRISPR-Cas: Adapting to change. *Science*, **356**, eaal5056.

56. Leenay, R.T. and Beisel, C.L. (2017) Deciphering, communicating, and engineering the CRISPR PAM. *J Mol Biol*, **429**, 177-191.

57. Horvath, P., Romero, D.A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol*, **190**, 1401-1412.

58. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol*, **190**, 1390-1400.

59. Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S. *et al.* (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, **520**, 186-191.

60. Shah, S.A., Erdmann, S., Mojica, F.J.M. and Garrett, R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol*, **10**, 891-899.

61. Yin, Y., Yang, B. and Entwistle, S. (2019) Bioinformatics identification of anti-CRISPR loci by using homology, guilt-by-association, and CRISPR self-targeting spacer approaches. *mSystems*, **4**, e00455-00419.

62. Pawluk, A., Staals, R.H.J., Taylor, C., Watson, B.N.J., Saha, S., Fineran, P.C., Maxwell, K.L. and Davidson, A.R. (2016) Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat Microbiol*, **1**, 16085.

63. Marino, N.D., Zhang, J.Y., Borges, A.L., Sousa, A.A., Leon, L.M., Rauch, B.J., Walton, R.T., Berry, J.D., Joung, J.K., Kleinstiver, B.P. *et al.* (2018) Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science*, **362**, 240-242.

64. Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K. *et al.* (2019) Detection of spacer precursors formed in vivo during primed CRISPR adaptation. *Nat Commun*, **10**, 4603.

65. Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M. and Brouns, S.J.J. (2016) Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. *Mol Cell*, **63**, 852-864.

66. Li, M., Gong, L., Zhao, D., Zhou, J. and Xiang, H. (2017) The spacer size of I-B CRISPR is modulated by the terminal sequence of the protospacer. *Nucleic Acids Res*, **45**, 4642-4654.

67. Jackson, S.A., Birkholz, N., Malone, L.M. and Fineran, P.C. (2019) Imprecise spacer acquisition generates CRISPR-Cas immune diversity through primed adaptation. *Cell Host Microbe*, **25**, 250-260.e254.

68. Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A., Logacheva, M.D. and Severinov, K. (2016) Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc Natl Acad Sci U S A*, **113**, 7626-7631.

69. Pyenson, N.C. and Marraffini, L.A. (2019) Expansion of CRISPR loci with multiple memories of infection enables the survival of structured bacterial communities. *bioRxiv*, 747212.

70. Edgar, R. and Qimron, U. (2010) The Escherichia coli CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *J Bacteriol*, **192**, 6291-6294.

71. Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter, C., Przybilski, R., Pitman, A.R. and Fineran, P.C. (2013) Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet*, **9**, e1003454-e1003454.

72. Dolan, A.E., Hou, Z., Xiao, Y., Gramelspacher, M.J., Heo, J., Howden, S.E., Freddolino, P.L., Ke, A. and Zhang, Y. (2019) Introducing a spectrum of long-range genomic deletions in human embryonic stem cells using Type I CRISPR-Cas. *Mol Cell*, **74**, 936-950.e935.

73. Goldberg, G.W., Jiang, W., Bikard, D. and Marraffini, L.A. (2014) Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature*, **514**, 633-637.

74. O'Meara, D. and Nunney, L. (2019) A phylogenetic test of the role of CRISPR-Cas in limiting plasmid acquisition and prophage integration in bacteria. *Plasmid*, **104**, 102418.

75. Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R. and Marraffini, L.A. (2013) Dealing with the evolutionary downside of CRISPR immunity: Bacteria and beneficial plasmids. *PLOS Genet*, **9**, e1003844.

76. Bikard, D., Hatoum-Aslan, A., Mucida, D. and Marraffini, Luciano A. (2012) CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe*, **12**, 177-186.

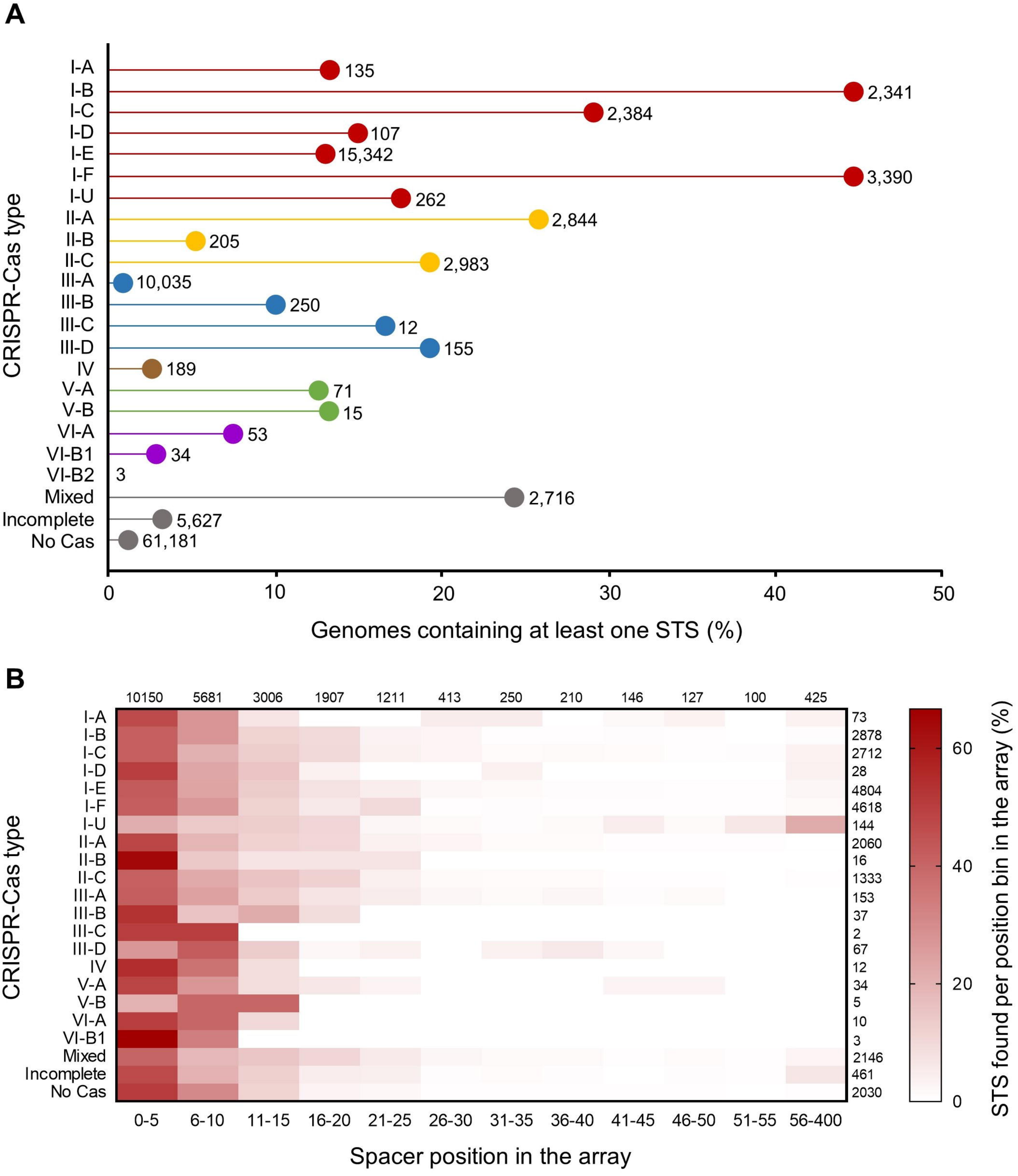
FIGURE LEGENDS

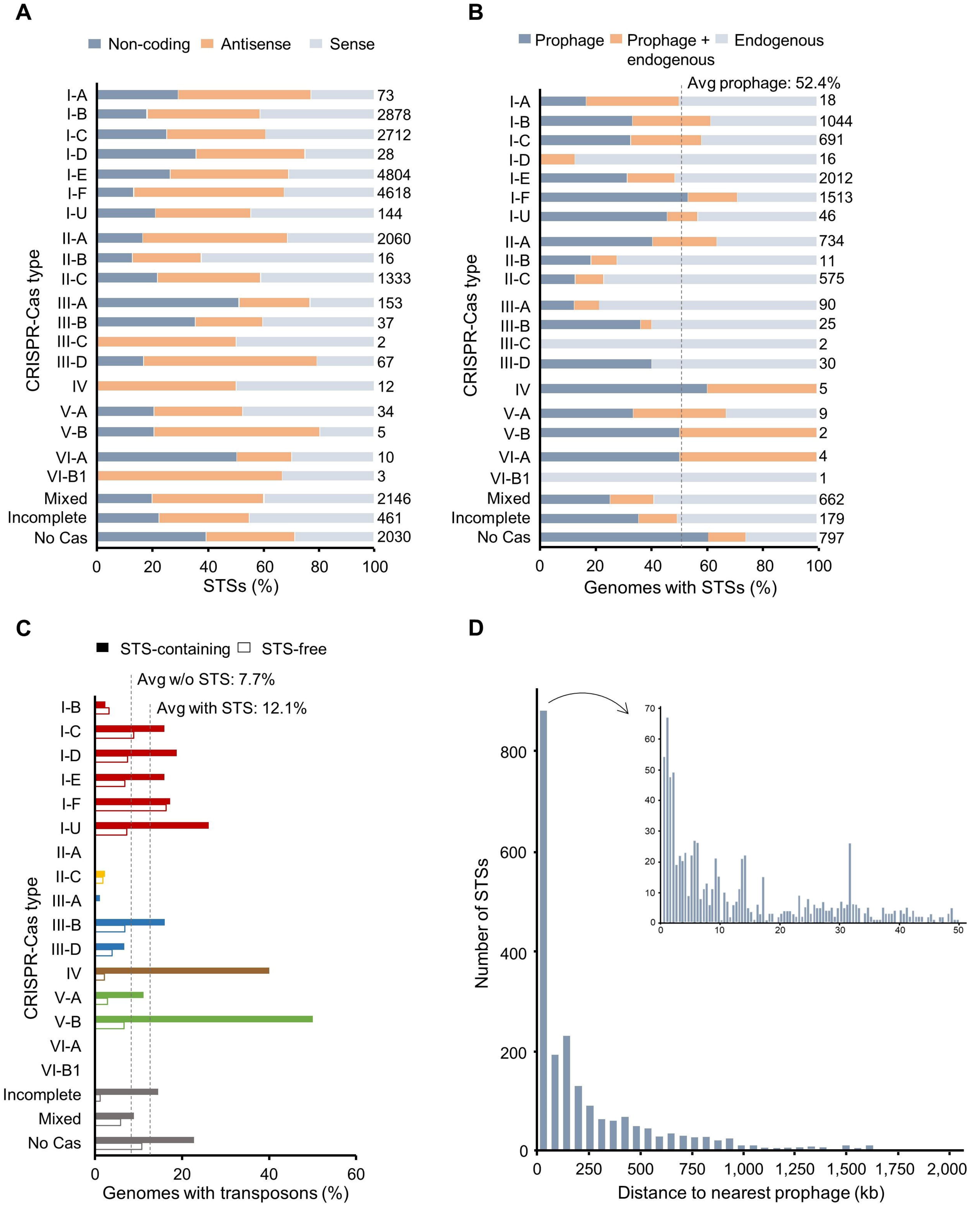
Figure 1. Self-targeting spacers (STS) in CRISPR-containing bacteria. (**A**) Frequency of genomes containing STS for the different subtypes of CRISPR-Cas systems. Total number of CRISPR-containing genomes analyzed is given for each row; (**B**) Heatmap of STS position in the CRISPR array for each CRISPR-Cas subtype, using corrected orientation of the CRISPR arrays. Scale bar represents percentage of STS found per position bin in the CRISPR array. Total number of STS analyzed per CRISPR-Cas subtype is given for each row, while total number of STS per position bin is given for each column.

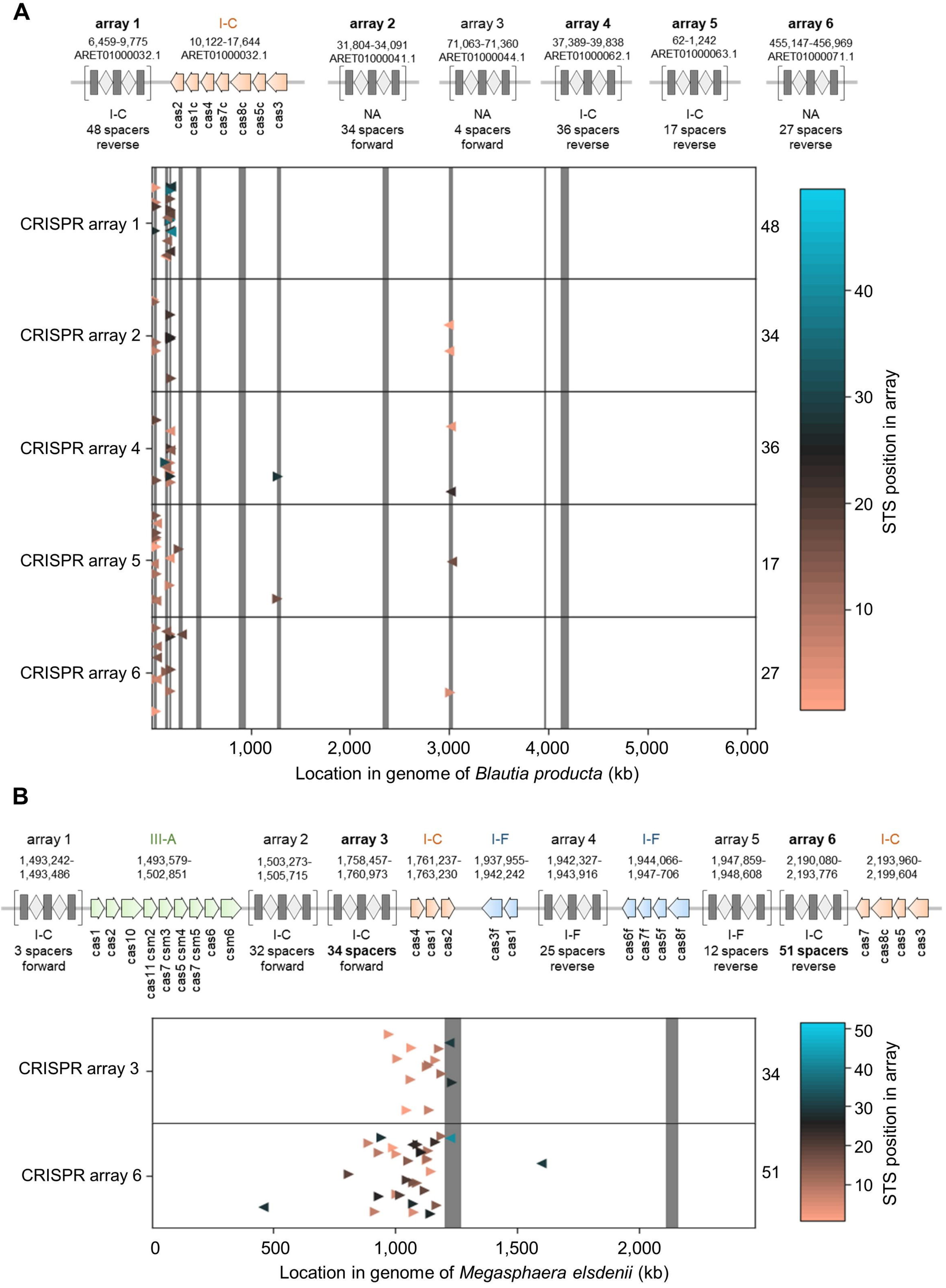
Figure 2. Genomic targets of self-targeting spacers (STS). (**A**) Preference of STS for targeting sense or antisense strands of coding regions, or non-coding regions of the bacterial genome. Values were normalized to the percentage of coding or non-coding regions of the genome. Total number of STS are indicated at the end of bars; (**B**) Prevalence of STS targeting only prophage regions, endogenous genomic regions, or both, in each CRISPR-Cas subtype. Total number of STS-containing genomes are indicated for bars; (**C**) Transposon abundance in STS-containing genomes (full bars) and STS-free genomes (empty bars) for each CRISPR-Cas subtype; (**D**) Distribution of distances between STS protospacer and the nearest prophage. Internal plot shows the largest peak binned into smaller (0.5 kb) increments.

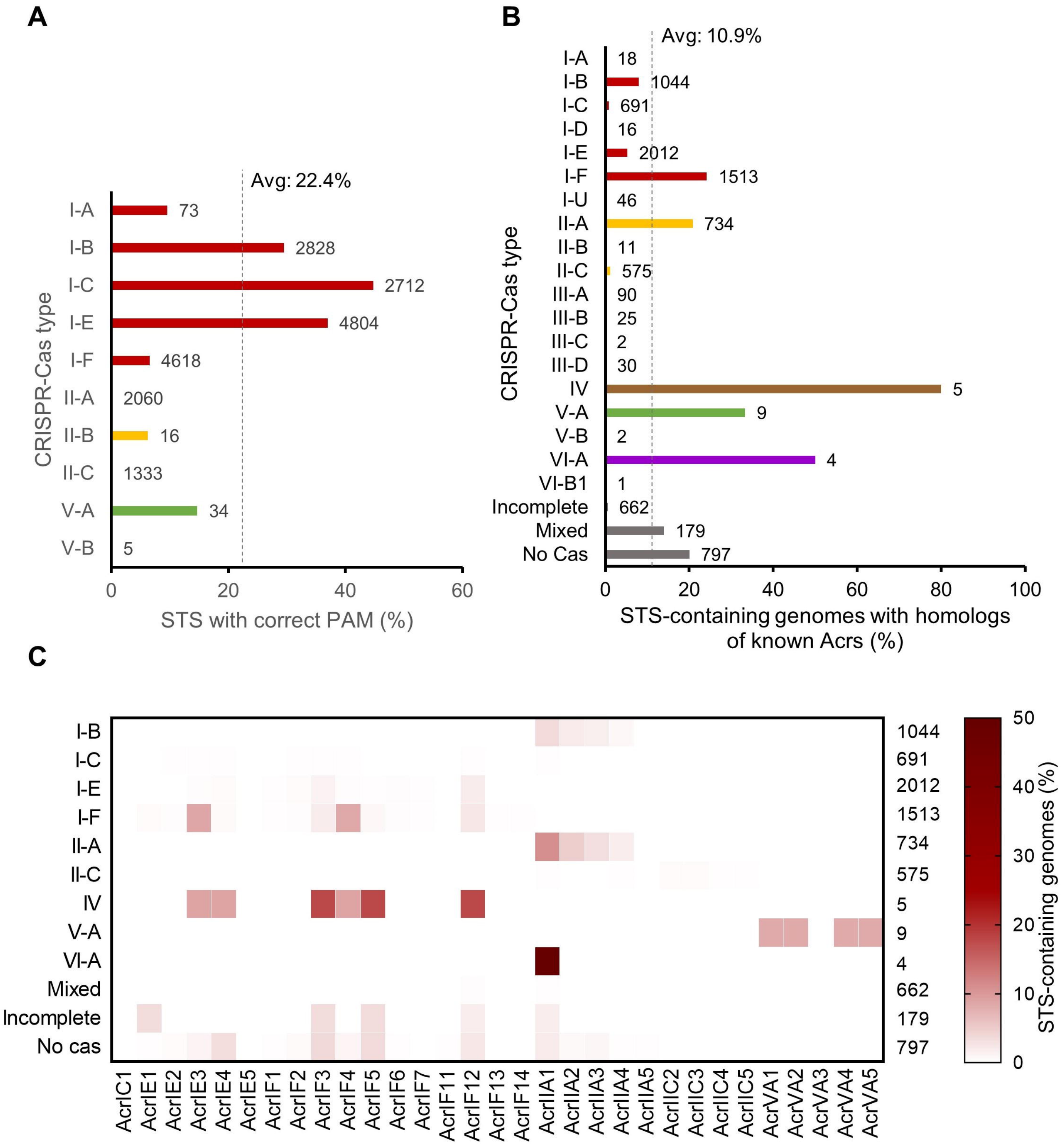
Figure 3. Extreme cases of self-targeting in prophage regions of bacterial genomes containing a high number of STS with 100% sequence identity to the target. (**A**) *Blautia producta* strain ATCC 27340 (accession number ARET01000032) carries a type I-C CRISPR-Cas system and 11 prophages, and has 162 STS. Arrays identified in different contigs from where STS originate are represented in the y-axis; and (**B**) *Megasphaera elsdenii* strain DSM 20460 (accession number NC_015873) carries types I-C, I-F and III-A CRISPR-Cas systems and two prophages, and has 85 STS. STS originate from two out of six CRISPR arrays (array 3 at 1,758,457-1,760,973 bp, and array 6 at 2,190,080-2,193,776 bp), which are associated with the type I-C system and are represented in the y-axis. For both panels, prophage regions are denoted in dark grey, STS hits are represented as colored triangles, and scale represents position of STS in the array. The total number of STS per contig or array is shown for each row.

Figure 4. Mechanisms of escape from auto-immunity. (**A**) Levels of self-targeting spacers (STS) associated with correct or incorrect protospacer adjacent motif (PAM) for different types of CRISPR-Cas systems. Only CRISPR-Cas systems with unquestionable type classification and of known PAM were considered. Dashed line indicates the average percentage of STS-containing genomes with correct PAM across CRISPR types; (**B**) Prevalence of STS-containing genomes with Acrs, as found by homology search to known Acrs. Dashed line indicates the average percentage of STS-containing genomes with Acr across CRISPR types; (**C**) Heatmap of prevalence of Acr families in different types of CRISPR-Cas systems in STS-containing genomes. Scale bar represents percentage of STS-containing genomes with a given CRISPR type (row) that contained a homolog of the Acr (column). The total number of STS-containing genomes of each CRISPR-Cas type is given at the end of each row.


A
CRISPR-Cas type
I-A 135
I-B 2,341
I-C 2,384
I-D 107
I-E 15,342
I-F 3,390
I-U 262
II-A 2,844
II-B 205
II-C 2,983
III-A 10,035
III-B 250
III-C 12
III-D 155
IV 189
V-A 71
V-B 15
VI-A 53
VI-B1 34
VI-B2 3
Mixed 2,716
Incomplete 5,627
No Cas 61,181
0
10
20
30
40
50
Genomes containing at least one STS (%)
B
10150
5681
3006
1907
1211
413
250
210
146
127
100
425
CRISPR-Cas type
I-A 73
I-B 2878
I-C 2712
I-D 28
I-E 4804
I-F 4618
I-U 144
II-A 2060
II-B 16
II-C 1333
III-A 153
III-B 37
III-C 2
III-D 67
IV 12
V-A 34
V-B 5
VI-A 10
VI-B1 3
Mixed 2146
Incomplete 461
No Cas 2030
0-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-50
51-55
56-400
Spacer position in the array
60
40
20
0
STS found per position bin in the array (%)


A
■ Non-coding ■ Antisense ■ Sense
CRISPR-Cas type
I-A 73
I-B 2878
I-C 2712
I-D 28
I-E 4804
I-F 4618
I-U 144
II-A 2060
II-B 16
II-C 1333
III-A 153
III-B 37
III-C 2
III-D 67
IV 12
V-A 34
V-B 5
VI-A 10
VI-B1 3
Mixed 2146
Incomplete 461
No Cas 2030
0 20 40 60 80 100
STSs (%)
B
■ Prophage ■ Prophage + endogenous ■ Endogenous
Avg prophage: 52.4%
CRISPR-Cas type
I-A 18
I-B 1044
I-C 691
I-D 16
I-E 2012
I-F 1513
I-U 46
II-A 734
II-B 11
II-C 575
III-A 90
III-B 25
III-C 2
III-D 30
IV 5
V-A 9
V-B 2
VI-A 4
VI-B1 1
Mixed 662
Incomplete 179
No Cas 797
0 20 40 60 80 100
Genomes with STSs (%)
C
■ STS-containing □ STS-free
Avg w/o STS: 7.7%
Avg with STS: 12.1%
CRISPR-Cas type
I-B
I-C
I-D
I-E
I-F
I-U
II-A
II-C
III-A
III-B
III-D
IV
V-A
V-B
VI-A
VI-B1
Incomplete
Mixed
No Cas
0 20 40 60
Genomes with transposons (%)
D
Number of STSs
0 200 400 600 800
0 250 500 750 1,000 1,250 1,500 1,750 2,000
Distance to nearest prophage (kb)
0 10 20 30 40 50 60 70
0 10 20 30 40 50


A
array 1
6,459-9,775
ARET01000032.1
I-C
48 spacers
reverse
I-C
10,122-17,644
ARET01000032.1
cas2
cas1c
cas4
cas7c
cas8c
cas5c
cas3
array 2
31,804-34,091
ARET01000041.1
NA
34 spacers
forward
array 3
71,063-71,360
ARET01000044.1
NA
4 spacers
forward
array 4
37,389-39,838
ARET01000062.1
I-C
36 spacers
reverse
array 5
62-1,242
ARET01000063.1
I-C
17 spacers
reverse
array 6
455,147-456,969
ARET01000071.1
NA
27 spacers
reverse
CRISPR array 1
48
CRISPR array 2
34
CRISPR array 4
36
CRISPR array 5
17
CRISPR array 6
27
1,000
2,000
3,000
4,000
5,000
6,000
Location in genome of Blautia producta (kb)
40
30
20
10
STS position in array
B
array 1
1,493,242-
1,493,486
I-C
3 spacers
forward
III-A
1,493,579-
1,502,851
cas1
cas2
cas10
cas11 csm2
cas7 csm3
cas5 csm4
cas7 csm5
cas6
csm6
array 2
1,503,273-
1,505,715
I-C
32 spacers
forward
array 3
1,758,457-
1,760,973
I-C
34 spacers
forward
I-C
1,761,237-
1,763,230
cas4
cas1
cas2
I-F
1,937,955-
1,942,242
cas3f
cas1
array 4
1,942,327-
1,943,916
I-F
25 spacers
reverse
I-F
1,944,066-
1,947-706
cas6f
cas7f
cas5f
cas8f
array 5
1,947,859-
1,948,608
I-F
12 spacers
reverse
array 6
2,190,080-
2,193,776
I-C
51 spacers
reverse
I-C
2,193,960-
2,199,604
cas7
cas8c
cas5
cas3
CRISPR array 3
34
CRISPR array 6
51
0
500
1,000
1,500
2,000
Location in genome of Megasphaera elsdenii (kb)
50
40
30
20
10
STS position in array


A
Avg: 22.4%
CRISPR-Cas type
I-A 73
I-B 2828
I-C 2712
I-E 4804
I-F 4618
II-A 2060
II-B 16
II-C 1333
V-A 34
V-B 5
0 20 40 60
STS with correct PAM (%)
B
Avg: 10.9%
CRISPR-Cas type
I-A 18
I-B 1044
I-C 691
I-D 16
I-E 2012
I-F 1513
I-U 46
II-A 734
II-B 11
II-C 575
III-A 90
III-B 25
III-C 2
III-D 30
IV 5
V-A 9
V-B 2
VI-A 4
VI-B1 1
Incomplete 662
Mixed 179
No Cas 797
0 20 40 60 80 100
STS-containing genomes with homologs of known Acrs (%)
C
I-B 1044
I-C 691
I-E 2012
I-F 1513
II-A 734
II-C 575
IV 5
V-A 9
VI-A 4
Mixed 662
Incomplete 179
No cas 797
AcrIC1 AcrIE1 AcrIE2 AcrIE3 AcrIE4 AcrIE5 AcrIF1 AcrIF2 AcrIF3 AcrIF4 AcrIF5 AcrIF6 AcrIF7 AcrIF11 AcrIF12 AcrIF13 AcrIF14 AcrIIA1 AcrIIA2 AcrIIA3 AcrIIA4 AcrIIA5 AcrIIC2 AcrIIC3 AcrIIC4 AcrIIC5 AcrVA1 AcrVA2 AcrVA3 AcrVA4 AcrVA5
50
40
30
20
10
0
STS-containing genomes (%)