bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 # Prophages are associated with extensive CRISPR-Cas auto-
2 # immunity

3

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

5

6 [1] Department of Bionanoscience, Delft University of Technology, Delft, Netherlands
7 [2] Kavli Institute of Nanoscience, Delft, Netherlands
8 [3] Theoretical Biology and Bioinformatics, Science4Life, Utrecht University, Utrecht,
9 Netherlands

10

11

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

14

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

17
18
19
20
21

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

46 **INTRODUCTION**

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

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

71 Here we take a closer look at STS in the many types and subtypes of CRISPR-Cas
72 systems to identify the incidence, distribution and mechanism of evasion of potential
73 CRISPR-Cas auto-immunity in bacteria. We demonstrate that STS are frequently
74 observed in bacterial genomes, and that bacteria have evolved mechanisms to evade
75 death by auto-immunity while preserving their CRISPR-Cas systems. We propose that the
76 integration of phages in the bacterial chromosome provides evolutionary advantages to the
77 bacteria (e.g. acquisition of virulence traits) but is also the primary trigger of STS
78 acquisition in CRISPR arrays. We further suggest that mechanisms of evasion from auto-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

79 immunity create tolerance to the integrated invaders, benefiting both bacteria and phage
80 populations by allowing the acquisition of novel genetic information by the bacteria, and by
81 promoting phage (passive) dissemination in the bacterial population.

82

83 Material and methods

84 Detection of CRISPR arrays

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

88 Detection of self-targeting spacers

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

102 Classification of CRISPR-Cas systems

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

112 **Analysis of the genomic target**

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

127 **Distance between self-targeting spacer and prophages**

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

130 **Identification of anti-CRISPR proteins**

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

133 **Statistical analysis**

134 A binomial test was performed on CRISPR arrays of different sizes to test the hypothesis
135 that STS at the leader side of the CRISPR array are more common. Only STS from
136 CRISPR arrays smaller than 50 spacers were considered because larger arrays are too
137 scarce to result in a reliable statistical analysis. A chi-squared test was used to determine

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

138 statistical significance between percentages of populations. Statistical significance was
139 considered for $P < 0.05$.

140 **Software**

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

147

148 # Results

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

150 We scanned 43,526 CRISPR-encoding genomes for spacers with >90% sequence identity
151 to the endogenous genomic sequence that is not part of a CRISPR array. We decided
152 upon this definition of STS as a 10% mismatch between spacer and target can still trigger
153 a functional CRISPR response (direct interference and/or priming in type I) in many
154 CRISPR-Cas types (36-41). For clarity, we note that our definition of STS may exclude or
155 include certain sequences as a result. For example, STS protospacers that suffered
156 extensive mutations may be excluded, while spacers that target non-genomic regions of
157 high similarity to a genomic region may be included. We found that 23,626 out of
158 1,481,476 spacers (1.6%) are self-targeting based on this cutoff. Approximately half of
159 those (12,121, 0.8%) had 100% sequence identity to the genome from which the spacers
160 were derived (frequency of STS with mismatches can be seen in Supplementary Table
161 S4), a percentage higher than previously reported (0.4% with 100% identity) (14). Similar
162 to previous observations with smaller datasets (14), about one fifth (19%, 8,466) of
163 CRISPR-encoding genomes have at least one STS in one of their CRISPR arrays.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

170 It has been suggested that following the integration of an STS, the CRISPR-Cas system
171 must become inactivated in order to survive, and that this phenomenon could explain the
172 abundance of highly degraded CRISPR systems that contain *cas* pseudogenes (14).
173 Recent experimental evolution studies have shown that large genomic deletions
174 encompassing the entire CRISPR-Cas locus can occur as a consequence of auto-
175 immunity to prophages (45). We observed that 12% (979 of 8,466) of the STS-containing
176 genomes contain incomplete CRISPR systems or no *cas* genes, while 88% (7,490 of
177 8,466) seem to carry intact CRISPR-Cas systems ($P < 0.0001$, chi-squared t-test, Figure
178 1A). This suggests that CRISPR-Cas deletion can occur as a mechanism to survive STS,
179 but self-targeting can also be overcome through other mechanisms. To note that our
180 homology-based analysis cannot account for small inactivating mutations in *cas* genes that
181 could also render a CRISPR-Cas system non-functional, but we expect that the effect of
182 such recent pseudogenization is minor as inactive pseudogenes tend to be rapidly lost
183 from the genome (46,47). Moreover, we found that most STS locate in the leader proximal
184 positions of the array (Figure 1B, Supplementary Figure S2), with several STS also found
185 in middle and leader distal positions (Figure 1B). To account for potential bias introduced
186 in this analysis by smaller arrays, we generated the same plot for arrays of 10 or less
187 spacers (Supplementary Figure S3). The same trend is apparent, confirming that STS
188 preferably locate near the leader but are also present in later positions in the array. This
189 suggests that the CRISPR system (or at least memory acquisition) remains active after
190 integration of an STS into the CRISPR array and the cell remains viable. Correct CRISPR
191 array orientation prediction remains challenging in some cases (48), and there may be
192 some arrays in our database whose orientation was predicted incorrectly by the
193 CRISPRDirection tool. This may lead to noise in the positionality of STS. Still, we are
194 confident on our overall observations as CRISPRDirection is backed up by experimental
195 evidence for most CRISPR types, including type I-U (49).

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

200

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

201 **STS are enriched in prophage-containing genomes**

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

215 Bacteriophages are common targets of CRISPR-Cas systems and exist abundantly in
216 nature. Because some bacteriophages can integrate into the bacterial chromosome, we
217 next investigated if the presence of prophages in a genome would associate with the
218 presence of STS. We identified prophage regions in the STS-containing genomes and
219 observed that, on average, 52.4% of the STS-containing genomes have STS with
220 protospacers in prophage regions, with type I-F CRISPR-Cas systems showing up to 70%
221 genomes with prophage hits (Figure 2B). Interestingly, we also observed that 96.9%
222 (8,203 out of 8,466) of the STS-containing genomes have at least one integrated
223 prophage, while only 28.5% (9,992 out of 35,060) of the STS-free genomes contain
224 prophages (P < 0.0001, chi-squared test). It therefore appears that STS is linked to
225 carrying prophages.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

253

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

256 To explain how STS are tolerated we first looked at the targeting requirements of CRISPR-
257 Cas systems. In many CRISPR-Cas systems, the correct identification of the target is
258 dependent on a small 2-6 base pair motif immediately adjacent to the target DNA
259 sequence, known as the PAM (54). The PAM is essential for binding to and cleavage of
260 the target DNA by the Cas nucleases, and mutations in this sequence can abrogate
261 targeting (55). To understand how often STS protospacers have a consensus PAM, and
262 can therefore be efficiently targeted, we compared the PAM sequence of the STS
263 protospacer with the expected PAM sequence for the different CRISPR-Cas types

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

264 previously described (Supplementary Table S2) (24,25,56). We observed that 22.4% of all
265 STS (4,140 of 18,483 STS with 90% sequence identity) and 23.9% of STS with 100 %
266 identity (2,294 of 9,605) have a consensus PAM (Figure 4A and Supplementary Table S6),
267 suggesting these to be functional for direct interference. Type I CRISPR-Cas systems,
268 especially types I-B (29.5%), I-C (44.7%) and I-E (37.0%) have more STS with a
269 consensus PAM (average 27.5%) than type II (average 0.1%) or type V (average 12.8%)
270 (Figure 4A). This may suggest that bacteria encoding type II and type V systems avoid the
271 lethal effects of auto-immunity by having non-functional STS, while bacteria encoding type
272 I systems may employ other evasion mechanisms to withstand the lethal auto-immunity
273 effects of interference-functional STS.

274 Several factors should be considered when analyzing the role of PAM sequences in
275 tolerance mechanisms to STS. First, the full diversity of functional PAM sequences in
276 nature currently remains unknown, as does their distribution across taxa. Second, PAM
277 sequences can vary widely even within a CRISPR subtype (e.g. in different species)
278 (54,57-59). Third, different CRISPR class I (type I, III and IV) systems may use different
279 PAM sequences for spacer acquisition and for targeting (60). Our analysis has revealed a
280 range of candidate bacteria that can contain mechanisms allowing them to remain viable
281 while carrying interference-functional STS with known consensus PAM sequences. It will
282 be interesting to see these mechanisms further unraveled in future studies.

283

284 **Acrs are more prevalent in bacteria carrying STS**

285 To understand how bacteria are able to survive STS while keeping their *cas* genes intact,
286 we assessed the presence of Acrs encoded by prophages. By inhibiting the activity of the
287 CRISPR-Cas system using a variety of mechanisms (reviewed in (29)), Acrs can prevent
288 the lethal effects of STS auto-immunity. In fact, STS have been used to identify new Acr
289 proteins (13,61). We mapped Acrs in the STS-containing genomes using homology
290 searches with all currently known Acrs (29). Acrs were found at low frequency (10.9%
291 average, Figure 4B) but still at levels significantly higher than those found in STS-free,
292 CRISPR-containing genomes (0.3% average, $P < 0.0001$, chi-squared test). The levels of
293 Acrs here reported are a lower bound, as unidentified Acrs may be present in these
294 genomes and these proteins may thus have a higher influence in escaping auto-immunity.
295 Even so, we found many Acr homologs in STS-containing bacteria carrying single type I-B,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

296 IV or VI-A CRISPR-Cas systems, for which no Acrs have yet been described (Figure 4B
297 and Supplementary Table S7). Putative Acrs for type I-B and type IV CRISPR-Cas
298 systems were recently identified by using a bioinformatics pipeline (61), but to our
299 knowledge none has yet been suggested for type VI-A.

300 Among the newly found Acrs, homologs of AcrIF2-7, AcrIF11-13 and AcrIIA1-4 were the
301 most common in STS-containing genomes (Figure 4C). Interestingly, homologs of AcrIF1-
302 14, AcrIE1-5, and AcrIIA1-4 were found in genomes of diverse CRISPR-Cas subtypes,
303 while homologs of AcrVA1-5 and AcrIIC2-5 appear only in genomes containing the
304 corresponding CRISPR-Cas subtype. Particularly, homologs of AcrIF1-14 and AcrIE1-5
305 were found in type I and type IV CRISPR-Cas types, while homologs of AcrIIA1-4 were
306 detected in type I, II and VI-A CRISPR-Cas systems. Acr homologs of families that do not
307 correspond to the CRISPR-Cas system found in the bacteria were also recently reported
308 (61). It is possible that some Acr homologs have activity against multiple types of CRISPR-
309 Cas systems, which may occur if the mechanism of inhibition of the Acr is compatible with
310 the multiple types. The ability of Acrs to inhibit different types of CRISPR-Cas systems has
311 already been revealed for some Acrs (62,63), although the specific mechanisms of
312 inhibition have not yet been described.

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

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

319

320 **Amplified self-targeting in prophages regions**

321 In our analysis, we found 1,224 genomes with a number of STS higher than the average
322 (2.5 ± 2.9 STS, Supplementary Figure S5). We decided to take a closer look at two
323 extreme cases and investigate how STS with 100% identity were distributed in the
324 bacterial chromosome (Figure 3). The genome of *Blautia producta* strain ATCC 27340
325 contains a type I-C CRISPR-Cas system and 11 prophage regions in the chromosome
326 (Figure 3A). This strain contains a stunning 162 STS mostly hitting prophage regions. The

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

327 genome of *Megasphaera elsdenii* strain DSM 20460 contains three distinct CRISPR-Cas
328 systems (types I-C, I-F and III-A), two large prophage regions (Figure 3B) and a total of 85
329 STS in its I-C CRISPR arrays. In *B. producta* and *M. elsdenii*, the wealth of STS hit mostly
330 in and around prophage regions, with some prophages remaining untargeted. After
331 manual confirmation of the consensus repeat and array orientation of the STS, we
332 observed that the oldest STS (located further from the leader in the CRISPR array) are
333 those with protospacer in the prophage regions (Figure 3A and 3B), suggesting these were
334 the initial hits and that additional spacers could have been acquired from locations in the
335 prophage vicinity by primed CRISPR adaptation. Interestingly, as priming is enhanced by
336 CRISPR interference (18,64,65), it is striking that no apparent DNA damage was incurred.
337 For *M. elsdenii* we found that all STS protospacers are on the same strand with an
338 orientation bias characteristic of primed adaptation (18). Primed adaptation would result in
339 the acquisition of many spacers, explaining the high number of STS found in these
340 genomes. It is interesting that STS in *M. elsdenii* were integrated in only two out of six
341 CRISPR arrays, both close to the I-C *cas* genes (Figure 3B). It is also curious to note that
342 no homologs of any known Acr (29) could be found in either genome using BLASTp
343 homology searches with an e-value cutoff of 10^{-5}.

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

347

348 **DISCUSSION**

349 Self-targeting CRISPR spacers (STS) in bacteria are not a rare phenomenon, as one fifth
350 of bacteria with CRISPR systems carries STS. Interestingly, some types of CRISPR-Cas
351 systems (i.e. types I-B and I-F) seem to be more prone to incorporation of STS into
352 CRISPR arrays. As STS may lead to auto-immunity, here we questioned which
353 mechanisms could drive STS acquisition and whether bacteria encode mechanisms to
354 protect themselves. We observed a striking prevalence of prophages in STS-containing
355 genomes when compared to STS-free genomes, suggesting that prophages could be the
356 trigger of STS acquisition. Only about half of the STS targeted protospacers are located
357 within the prophage regions, with the other half targeting the endogenous genome.
358 Interestingly, STS hits in the endogenous genome are enriched in the proximity of

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

359 prophages, showing a pattern consistent with primed adaptation from an initial protospacer
360 present on the prophage. Also, in cases where bacteria carried multiple STS, the STS
361 located the furthest from the leader sequence targeted the prophage, while subsequent
362 STS targeted both prophage and endogenous regions. These results are consistent with a
363 model where primed adaptation amplifies STS by acquisition of new spacers from both
364 prophage and prophage-adjacent regions.

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

378 Based on our overall observations, we here suggest two scenarios for the appearance of
379 STS in bacterial genomes. In the first scenario, bacteria may acquire a first spacer against
380 a temperate phage, but despite this, the phage may still be able to integrate into the
381 genome. In the second scenario, a prophage may already be integrated into the genome
382 and the ‘accidental’ acquisition of an STS by the host may start targeting the prophage.
383 Following this first STS, incomplete targeting may lead to further STS expansion by primed
384 spacer acquisition, in type I and II systems (68,69), which will result in the incorporation of
385 multiple new spacers targeting both the prophage and adjacent locations in the bacterial
386 genome. This continuation of extensive priming, which is thought to require a level of
387 CRISPR targeting, is without apparent genome damage or lethality. The process of
388 acquiring STS creates an apparent standoff between CRISPR-Cas and targeted
389 prophages that involves mechanisms of auto-immunity avoidance and anti-phage defense.
390 As shown here, these interactions may involve Acrs that may contribute to creating
391 tolerance to STS in general, and to STS-targeted prophages in particular. Thus, it is

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

392 possible that the CRISPR system may be preventing prophage induction (70), and
393 perhaps induce prophage clearance or genome deletions (71,72). When the protospacer
394 region of the prophage in the bacterial genome is deleted, this may lead to interesting eco-
395 evolutionary dynamics, as the presence of the former STS on the bacterial genome may
396 now prevent reinfection of the immunized strain by the same or related phages. Similarly, if
397 the CRISPR system prevents induction of the prophage by targeting it upon excision from
398 the genome, the induction of the lytic cycle could be inhibited and the shift from lysogeny
399 to a lytic state could be detected and acted upon. The balance between these processes
400 remains subject to further experimentation and modelling.

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

414

415 **AVAILABILITY**

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

418

419 **SUPPLEMENTARY DATA**

420 Supplementary Data are available at NAR online.

421

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

422 **FUNDING**

423 This work was supported by the Netherlands Organisation for Scientific Research (NWO)
424 [grant numbers VENI 016.Veni.181.092 to F.L.N., VIDI 864.14.004 to B.E.D., and VICI
425 VI.C.182.027 to S.J.J.B.]; and European Research Council (ERC) Stg grant [grant number
426 639707 to S.J.J.B.] and Consolidator grant [grant number 865694 to B.E.D]. Funding for
427 open access charge: NWO [grant number 016.Veni.181.092].

428

429 **CONFLICT OF INTEREST**

430 None declared.

431

432 **REFERENCES**

433 1. Jansen, R., Embden, J.D.A.v., Gaastra, W. and Schouls, L.M. (2002) Identification of
434 genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565-
435 1575.
436 2. Mojica, F.J.M., Díez-Villaseñor, C.s., García-Martínez, J. and Soria, E. (2005)
437 Intervening sequences of regularly spaced prokaryotic repeats derive from foreign
438 genetic elements. *J Mol Evol*, **60**, 174-182.
439 3. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S.,
440 Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against
441 viruses in prokaryotes. *Science*, **315**, 1709-1712.
442 4. Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders,
443 A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V. and van der Oost, J. (2008)
444 Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960-964.
445 5. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. and
446 Terns, M.P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein
447 complex. *Cell*, **139**, 945-956.
448 6. Wimmer, F. and Beisel, C.L. (2020) CRISPR-Cas systems and the paradox of self-
449 targeting spacers. *Front Microbiol*, **10**, 3078-3078.
450 7. Heussler, G.E. and O'Toole, G.A. (2016) Friendly Fire: Biological functions and
451 consequences of chromosomal targeting by CRISPR-Cas systems. *J Bacteriol*, **198**,
452 1481-1486.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

453 8. Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V. and Koonin,
454 E.V. (2017) The CRISPR spacer space is dominated by sequences from species-
455 specific mobilomes. *mBio*, **8**, e01397-01317.
456 9. Selle, K., Klaenhammer, T.R. and Barrangou, R. (2015) CRISPR-based screening of
457 genomic island excision events in bacteria. *Proc Natl Acad Sci U S A*, **112**, 8076-
458 8081.
459 10. Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., Chen, L., Liang, Y.X. and She,
460 Q. (2016) Harnessing Type I and Type III CRISPR-Cas systems for genome editing.
461 *Nucleic Acids Res*, **44**, e34-e34.
462 11. Rauch, B.J., Silvis, M.R., Hultquist, J.F., Waters, C.S., McGregor, M.J., Krogan, N.J.
463 and Bondy-Denomy, J. (2017) Inhibition of CRISPR-Cas9 with bacteriophage
464 proteins. *Cell*, **168**, 150-158.e110.
465 12. Bondy-Denomy, J. (2018) Protein inhibitors of CRISPR-Cas9. *ACS Chem Biol*, **13**,
466 417-423.
467 13. Watters, K.E., Fellmann, C., Bai, H.B., Ren, S.M. and Doudna, J.A. (2018) Systematic
468 discovery of natural CRISPR-Cas12a inhibitors. *Science*, **362**, 236-239.
469 14. Stern, A., Keren, L., Wurtzel, O., Amitai, G. and Sorek, R. (2010) Self-targeting by
470 CRISPR: gene regulation or autoimmunity? *Trends Genet*, **26**, 335-340.
471 15. Cady, K.C. and O'Toole, G.A. (2011) Non-identity-mediated CRISPR-bacteriophage
472 interaction mediated via the Csy and Cas3 proteins. *J Bacteriol*, **193**, 3433-3445.
473 16. Li, R., Fang, L., Tan, S., Yu, M., Li, X., He, S., Wei, Y., Li, G., Jiang, J. and Wu, M.
474 (2016) Type I CRISPR-Cas targets endogenous genes and regulates virulence to
475 evade mammalian host immunity. *Cell Res*, **26**, 1273-1287.
476 17. Ratner, H.K., Escalera-Maurer, A., Le Rhun, A., Jaggavarapu, S., Wozniak, J.E.,
477 Crispell, E.K., Charpentier, E. and Weiss, D.S. (2019) Catalytically active Cas9
478 mediates transcriptional interference to facilitate bacterial virulence. *Mol Cell*, **75**, 498-
479 510.e495.
480 18. Staals, R.H.J., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M. and Fineran,
481 P.C. (2016) Interference-driven spacer acquisition is dominant over naive and primed
482 adaptation in a native CRISPR-Cas system. *Nat Commun*, **7**, 12853-12853.
483 19. Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N.,
484 Dietrich, E.M., Disz, T., Gabbard, J.L. *et al.* (2016) Improvements to PATRIC, the all-
485 bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res*,
486 **45**, D535-D542.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

487 20. Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C. and Brown, C.M. (2016)
488 CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, **17**,
489 356-356.
490 21. Abby, S.S., Néron, B., Ménager, H., Touchon, M. and Rocha, E.P.C. (2014)
491 MacSyFinder: A program to mine genomes for molecular systems with an application
492 to CRISPR-Cas systems. *PLoS ONE*, **9**, e110726.
493 22. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J.
494 (2010) Prodigal: prokaryotic gene recognition and translation initiation site
495 identification. *BMC Bioinformatics*, **11**, 119.
496 23. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B.,
497 Rocha, E.P.C., Vergnaud, G., Gautheret, D. and Pourcel, C. (2018)
498 CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced
499 performance and integrates search for Cas proteins. *Nucleic Acids Res*, **46**, W246-
500 W251.
501 24. Westra, E.R., Swarts, D.C., Staals, R.H.J., Jore, M.M., Brouns, S.J.J. and Oost, J.v.d.
502 (2012) The CRISPRs, They are a-changin': How Prokaryotes Generate Adaptive
503 Immunity. *Annu Rev Genet*, **46**, 311-339.
504 25. Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G. and
505 Randau, L. (2019) PAM identification by CRISPR-Cas effector complexes: diversified
506 mechanisms and structures. *RNA Biol*, **16**, 504-517.
507 26. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,
508 Maslen, J., Mitchell, A., Nuka, G. *et al.* (2014) InterProScan 5: genome-scale protein
509 function classification. *Bioinformatics*, **30**, 1236-1240.
510 27. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger,
511 A., Hetherington, K., Holm, L., Mistry, J. *et al.* (2014) Pfam: the protein families
512 database. *Nucleic Acids Res*, **42**, D222-D230.
513 28. Roux, S., Enault, F., Hurwitz, B. and Sullivan, M. (2015) VirSorter: mining viral signal
514 from microbial genomic data. *PeerJ*, **3**, e985.
515 29. Trasanidou, D., Gerós, A.S., Mohanraju, P., Nieuwenweg, A.C., Nobrega, F.L. and
516 Staals, R.H.J. (2019) Keeping crispr in check: diverse mechanisms of phage-encoded
517 anti-crisprs. *FEMS Microbiol Lett*, **366**, fnz098.
518 30. Tange, O. (2011) GNU Parallel - The Command-Line Power Tool. *The USENIX*
519 *Magazine*, **February 2011**, 42-47.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

520 31. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg,
521 I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available
522 Python tools for computational molecular biology and bioinformatics. *Bioinformatics*,
523 **25**, 1422-1423.
524 32. McKinney, W. (2010) Data structures for statistical computing in Python. *Proc of the*
525 *9th Python in Science Conf (SCIPY 2010)*, 51-56.
526 33. Jones, E., Oliphant, T. and Peterson, P. (2001) *SciPy: Open Source Scientific Tools*
527 *for Python*.
528 34. Hunter, J.D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science &*
529 *Engineering*, **9**, 90-95.
530 35. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S.,
531 Gemperline, D., Augspurger, T., Halchenko, Y., Cole, J. *et al.* (2018)
532 mwaskom/seaborn: v0.9.0 (July 2018). *Zenodo*,
533 http://doi.org/10.5281/zenodo.1313201.
534 36. Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., Boekhorst, J., van
535 Hijum, S.A.F.T., Staals, R.H.J. and Brouns, S.J.J. (2014) Degenerate target sites
536 mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A*, **111**, E1629-
537 E1638.
538 37. Goldberg, G.W., McMillan, E.A., Varble, A., Modell, J.W., Samai, P., Jiang, W. and
539 Marraffini, L.A. (2018) Incomplete prophage tolerance by type III-A CRISPR-Cas
540 systems reduces the fitness of lysogenic hosts. *Nat Commun*, **9**, 61-61.
541 38. Manica, A., Zebec, Z., Steinkellner, J. and Schleper, C. (2013) Unexpectedly broad
542 target recognition of the CRISPR-mediated virus defence system in the archaeon
543 Sulfolobus solfataricus. *Nucleic Acids Res*, **41**, 10509-10517.
544 39. Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B.,
545 van der Oost, J., Brouns, S.J.J. and Severinov, K. (2011) Interference by clustered
546 regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed
547 sequence. *Proc Natl Acad Sci U S A*, **108**, 10098-10103.
548 40. Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A. and Liu, D.R. (2013)
549 High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9
550 nuclease specificity. *Nat Biotechnol*, **31**, 839-843.
551 41. Zhang, B., Ye, Y., Ye, W., Perčulija, V., Jiang, H., Chen, Y., Li, Y., Chen, J., Lin, J.,
552 Wang, S. *et al.* (2019) Two HEPN domains dictate CRISPR RNA maturation and
553 target cleavage in Cas13d. *Nat Commun*, **10**, 2544.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

554 42. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. and Wang, Y. (2015)
555 Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-
556 Cas systems. *Cell*, **163**, 840-853.
557 43. Kuznedelov, K., Mekler, V., Lemak, S., Tokmina-Lukaszewska, M., Datsenko, K.A.,
558 Jain, I., Savitskaya, E., Mallon, J., Shmakov, S., Bothner, B. *et al.* (2016) Altered
559 stoichiometry Escherichia coli Cascade complexes with shortened CRISPR RNA
560 spacers are capable of interference and primed adaptation. *Nucleic Acids Res*, **44**,
561 10849-10861.
562 44. Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N. and Doudna, J.A.
563 (2015) Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, **527**,
564 535-538.
565 45. Rollie, C., Chevallereau, A., Watson, B.N.J., Chyou, T.-y., Fradet, O., McLeod, I.,
566 Fineran, P.C., Brown, C.M., Gandon, S. and Westra, E.R. (2020) Targeting of
567 temperate phages drives loss of type I CRISPR–Cas systems. *Nature*, **578**, 149-153.
568 46. Mira, A., Ochman, H. and Moran, N.A. (2001) Deletional bias and the evolution of
569 bacterial genomes. *Trends Genet*, **17**, 589-596.
570 47. Kuo, C.-H., Moran, N.A. and Ochman, H. (2009) The consequences of genetic drift for
571 bacterial genome complexity. *Genome Res*, **19**, 1450-1454.
572 48. Milicevic, O., Repac, J., Bozic, B., Djordjevic, M. and Djordjevic, M. (2019) A simple
573 criterion for inferring CRISPR array direction. *Front Microbiol*, **10**.
574 49. Almendros, C., Nobrega, F.L., McKenzie, R.E. and Brouns, S.J J. (2019) Cas4–Cas1
575 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids*
576 *Res*, **47**, 5223-5230.
577 50. Luo, M.L., Mullis, A.S., Leenay, R.T. and Beisel, C.L. (2015) Repurposing
578 endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic*
579 *Acids Res*, **43**, 674-681.
580 51. Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R. and Lundgren, M. (2015)
581 Efficient programmable gene silencing by Cascade. *Nucleic Acids Res*, **43**, 237-246.
582 52. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P. and
583 Lim, W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-
584 specific control of gene expression. *Cell*, **152**, 1173-1183.
585 53. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L.A. (2013)
586 Programmable repression and activation of bacterial gene expression using an
587 engineered CRISPR-Cas system. *Nucleic Acids Res*, **41**, 7429-7437.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

588 54. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J. and Almendros, C. (2009)
589 Short motif sequences determine the targets of the prokaryotic CRISPR defence
590 system. *Microbiology*, **155**, 733-740.
591 55. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C. and
592 Brouns, S.J.J. (2017) CRISPR-Cas: Adapting to change. *Science*, **356**, eaal5056.
593 56. Leenay, R.T. and Beisel, C.L. (2017) Deciphering, communicating, and engineering
594 the CRISPR PAM. *J Mol Biol*, **429**, 177-191.
595 57. Horvath, P., Romero, D.A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H.,
596 Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity,
597 and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol*, **190**, 1401-
598 1412.
599 58. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P.,
600 Romero, D.A., Horvath, P. and Moineau, S. (2008) Phage response to CRISPR-
601 encoded resistance in Streptococcus thermophilus. *J Bacteriol*, **190**, 1390-1400.
602 59. Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B.,
603 Shalem, O., Wu, X., Makarova, K.S. *et al.* (2015) In vivo genome editing using
604 Staphylococcus aureus Cas9. *Nature*, **520**, 186-191.
605 60. Shah, S.A., Erdmann, S., Mojica, F.J.M. and Garrett, R.A. (2013) Protospacer
606 recognition motifs: mixed identities and functional diversity. *RNA Biol*, **10**, 891-899.
607 61. Yin, Y., Yang, B. and Entwistle, S. (2019) Bioinformatics identification of anti-CRISPR
608 loci by using homology, guilt-by-association, and CRISPR self-targeting spacer
609 approaches. *mSystems*, **4**, e00455-00419.
610 62. Pawluk, A., Staals, R.H.J., Taylor, C., Watson, B.N.J., Saha, S., Fineran, P.C.,
611 Maxwell, K.L. and Davidson, A.R. (2016) Inactivation of CRISPR-Cas systems by anti-
612 CRISPR proteins in diverse bacterial species. *Nat Microbiol*, **1**, 16085.
613 63. Marino, N.D., Zhang, J.Y., Borges, A.L., Sousa, A.A., Leon, L.M., Rauch, B.J., Walton,
614 R.T., Berry, J.D., Joung, J.K., Kleinstiver, B.P. *et al.* (2018) Discovery of widespread
615 type I and type V CRISPR-Cas inhibitors. *Science*, **362**, 240-242.
616 64. Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I.,
617 Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K. *et al.* (2019)
618 Detection of spacer precursors formed in vivo during primed CRISPR adaptation. *Nat*
619 *Commun*, **10**, 4603.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

620 65. Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M.,
621 Depken, M., Suarez-Diez, M. and Brouns, S.J.J. (2016) Cas3-derived target DNA
622 degradation fragments fuel primed CRISPR adaptation. *Mol Cell*, **63**, 852-864.
623 66. Li, M., Gong, L., Zhao, D., Zhou, J. and Xiang, H. (2017) The spacer size of I-B
624 CRISPR is modulated by the terminal sequence of the protospacer. *Nucleic Acids*
625 *Res*, **45**, 4642-4654.
626 67. Jackson, S.A., Birkholz, N., Malone, L.M. and Fineran, P.C. (2019) Imprecise spacer
627 acquisition generates CRISPR-Cas immune diversity through primed adaptation. *Cell*
628 *Host Microbe*, **25**, 250-260.e254.
629 68. Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D.,
630 Datsenko, K.A., Logacheva, M.D. and Severinov, K. (2016) Highly efficient primed
631 spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-
632 Cas interfering complex. *Proc Natl Acad Sci U S A*, **113**, 7626-7631.
633 69. Pyenson, N.C. and Marraffini, L.A. (2019) Expansion of CRISPR loci with multiple
634 memories of infection enables the survival of structured bacterial communities.
635 *bioRxiv*, 747212.
636 70. Edgar, R. and Qimron, U. (2010) The Escherichia coli CRISPR system protects from λ
637 lysogenization, lysogens, and prophage induction. *J Bacteriol*, **192**, 6291-6294.
638 71. Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter,
639 C., Przybilski, R., Pitman, A.R. and Fineran, P.C. (2013) Cytotoxic chromosomal
640 targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or
641 remodel pathogenicity islands. *PLoS Genet*, **9**, e1003454-e1003454.
642 72. Dolan, A.E., Hou, Z., Xiao, Y., Gramelspacher, M.J., Heo, J., Howden, S.E.,
643 Freddolino, P.L., Ke, A. and Zhang, Y. (2019) Introducing a spectrum of long-range
644 genomic deletions in human embryonic stem cells using Type I CRISPR-Cas. *Mol*
645 *Cell*, **74**, 936-950.e935.
646 73. Goldberg, G.W., Jiang, W., Bikard, D. and Marraffini, L.A. (2014) Conditional tolerance
647 of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature*, **514**,
648 633-637.
649 74. O'Meara, D. and Nunney, L. (2019) A phylogenetic test of the role of CRISPR-Cas in
650 limiting plasmid acquisition and prophage integration in bacteria. *Plasmid*, **104**,
651 102418.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

652 75. Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R. and Marraffini, L.A. (2013)
653 Dealing with the evolutionary downside of CRISPR immunity: Bacteria and beneficial
654 plasmids. *PLOS Genet*, **9**, e1003844.
655 76. Bikard, D., Hatoum-Aslan, A., Mucida, D. and Marraffini, Luciano A. (2012) CRISPR
656 interference can prevent natural transformation and virulence acquisition during in vivo
657 bacterial infection. *Cell Host Microbe*, **12**, 177-186.

658

659

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

660 **FIGURE LEGENDS**

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

668

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

679

680 **Figure 3.** Extreme cases of self-targeting in prophage regions of bacterial genomes
681 containing a high number of STS with 100% sequence identity to the target. (**A**) *Blautia*
682 *producta* strain ATCC 27340 (accession number ARET01000032) carries a type I-C
683 CRISPR-Cas system and 11 prophages, and has 162 STS. Arrays identified in different
684 contigs from where STS originate are represented in the y-axis; and (**B**) *Megasphaera*
685 *elsdenii* strain DSM 20460 (accession number NC_015873) carries types I-C, I-F and III-A
686 CRISPR-Cas systems and two prophages, and has 85 STS. STS originate from two out of
687 six CRISPR arrays (array 3 at 1,758,457-1,760,973 bp, and array 6 at 2,190,080-
688 2,193,776 bp), which are associated with the type I-C system and are represented in the y-
689 axis. For both panels, prophage regions are denoted in dark grey, STS hits are
690 represented as colored triangles, and scale represents position of STS in the array. The
691 total number of STS per contig or array is shown for each row.

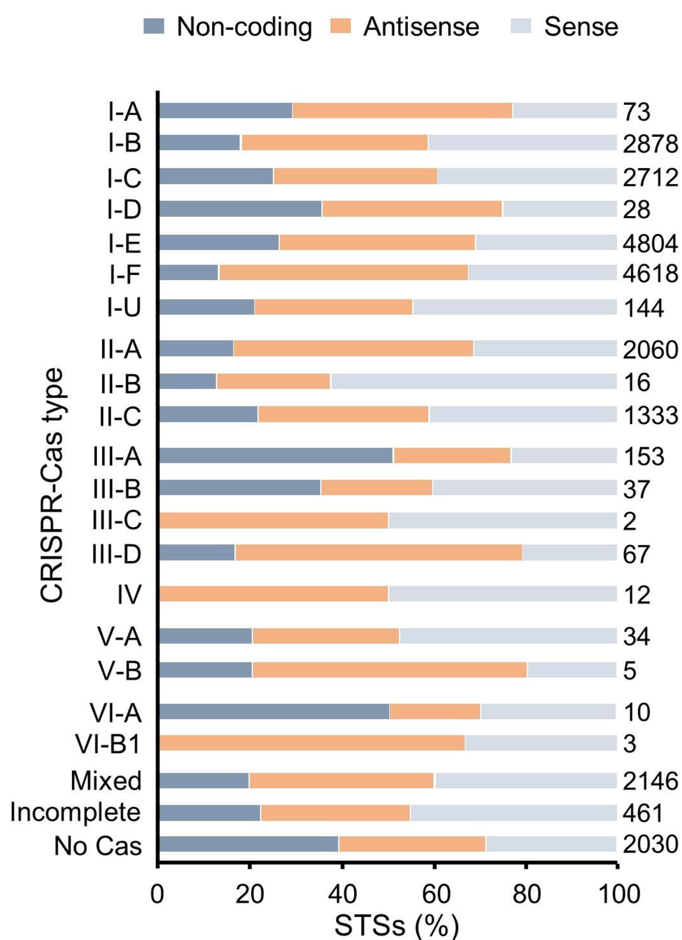
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.02.973784; this version posted August 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

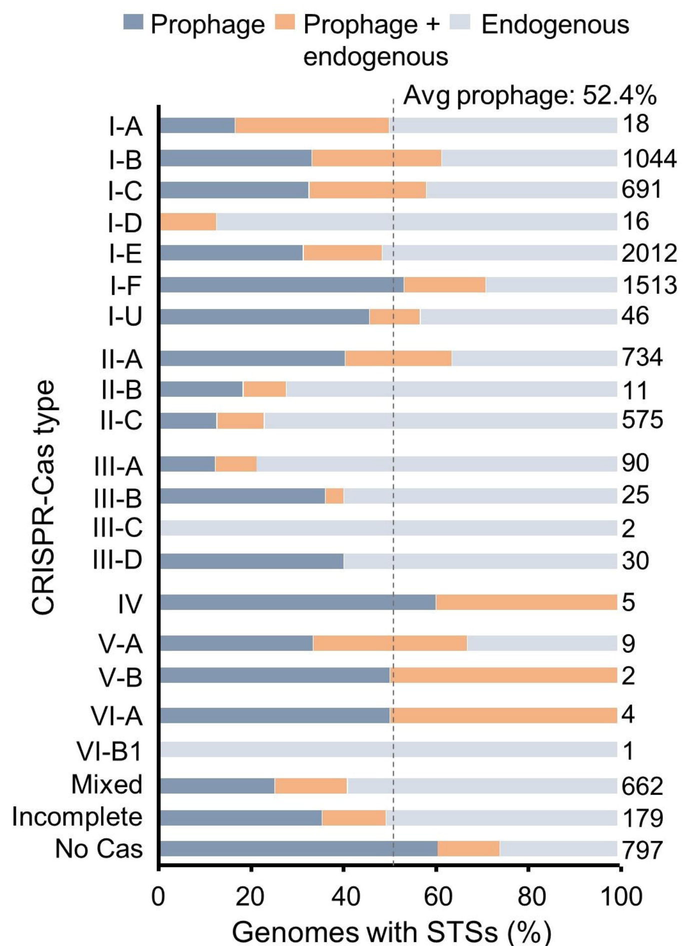
692

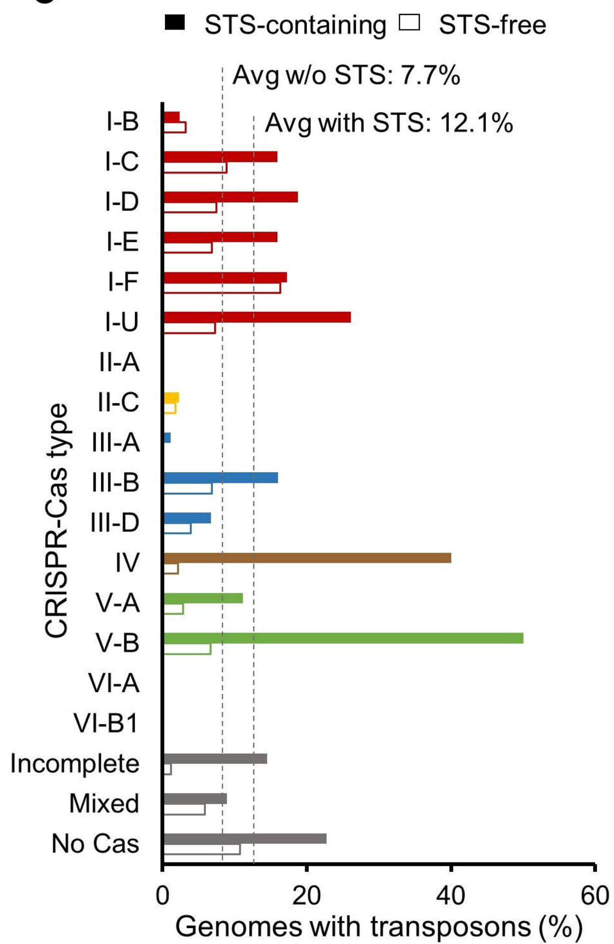
693 **Figure 4.** Mechanisms of escape from auto-immunity. (**A**) Levels of self-targeting spacers
694 (STS) associated with correct or incorrect protospacer adjacent motif (PAM) for different
695 types of CRISPR-Cas systems. Only CRISPR-Cas systems with unquestionable type
696 classification and of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

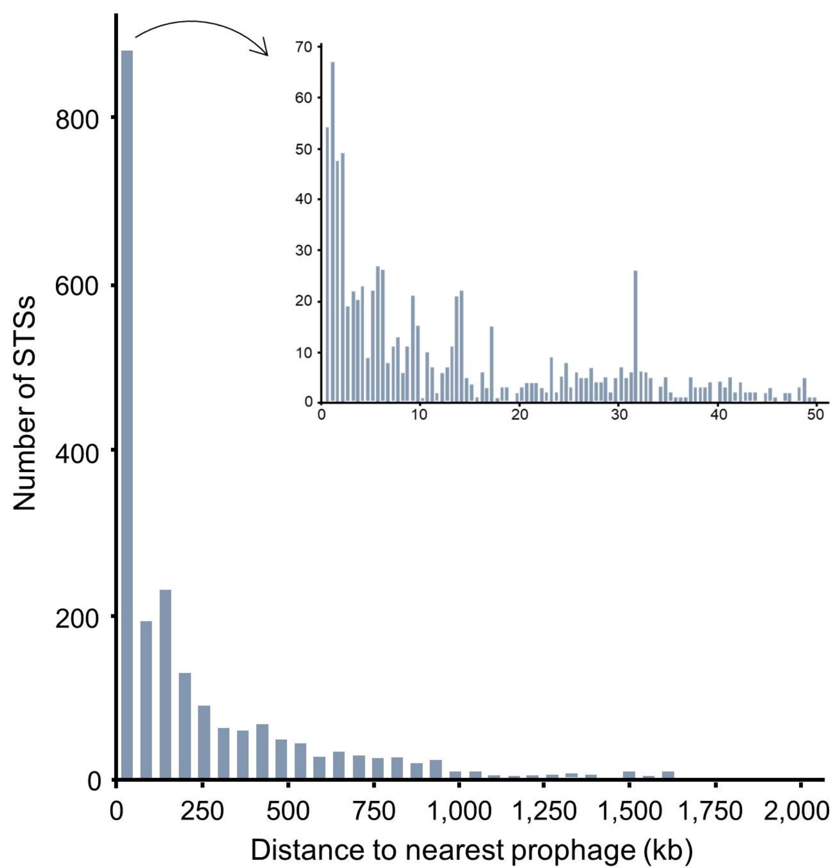
705

A
CRISPR-Cas type
Genomes containing at least one STS (%)
0
10
20
30
40
50
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
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
STS found per position bin in the array (%)
0
20
40
60


A
Non-coding
Antisense
Sense
I-A
I-B
I-C
I-D
I-E
I-F
I-U
II-A
II-B
II-C
III-A
III-B
III-C
III-D
IV
V-A
V-B
VI-A
VI-B1
Mixed
Incomplete
No Cas
73
2878
2712
28
4804
4618
144
2060
16
1333
153
37
2
67
12
34
5
10
3
2146
461
2030
CRISPR-Cas type
0
20
40
60
80
100
STSs (%)


B
Prophage
Prophage + endogenous
Endogenous
Avg prophage: 52.4%
I-A
I-B
I-C
I-D
I-E
I-F
I-U
II-A
II-B
II-C
III-A
III-B
III-C
III-D
IV
V-A
V-B
VI-A
VI-B1
Mixed
Incomplete
No Cas
18
1044
691
16
2012
1513
46
734
11
575
90
25
2
30
5
9
2
4
1
662
179
797
CRISPR-Cas type
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%
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
CRISPR-Cas type
0
20
40
60
Genomes with transposons (%)


D
800
600
400
200
0
Number of STSs
0
250
500
750
1,000
1,250
1,500
1,750
2,000
Distance to nearest prophage (kb)
70
60
50
40
30
20
10
0
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)

STS position in array
10 20 30 40

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)

STS position in array
10 20 30 40 50

A

Avg: 22.4%

CRISPR-Cas type

Type	Count
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

Type	Count
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

Type	Count
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

STS-containing genomes (%)

0 10 20 30 40 50