Is having more than one CRISPR array adaptive?

Jake L. Weissman[1], William F. Fagan[1], Philip L.F. Johnson [1*]

1 Department of Biology, University of Maryland College Park, MD, USA

* plfj@umd.edu

Abstract

Prokaryotes are under nearly constant attack by viral pathogens. To protect against this threat of infection, bacteria and archaea have evolved a wide array of defense mechanisms, singly and in combination. While immune diversity in a single organism likely reduces the chance of pathogen evolutionary escape, it remains puzzling why many prokaryotes also have multiple, seemingly redundant, copies of the same type of immune system. Here, we focus on the highly flexible CRISPR adaptive immune system, which is present in multiple copies in a surprising 21% of the prokaryotic genomes in RefSeq. We use a comparative genomics approach looking across all prokaryotes to demonstrate that having more than one CRISPR system confers a selective advantage to the organism, on average. We hypothesize that a tradeoff between memory span and learning speed could select for both "long-term memory" and "short-term memory" CRISPR arrays, and we go on to develop a mathematical model to confirm that such a tradeoff could lead to selection for multiple arrays.

Introduction

Just as larger organisms must cope with the constant threat of infection by pathogens, so too must bacteria and archaea. To defend themselves in a given pathogenic environment, prokaryotes may employ a range of different defense mechanisms, and oftentimes more than one [1–3]. While having multiple types of immune systems may decrease the chance of pathogen evolutionary escape [4], having multiple instances of the same type of system is rather more puzzling. Why have more than one of the same type of immune system? Here we endeavor to answer this question in the context of CRISPR-Cas immunity.

The CRISPR-Cas immune system is a powerful defense mechanism against the viruses that infect bacteria and archaea and is the only example of adaptive immunity in prokaryotes [5, 6]. This system allows prokaryotes to acquire specific immune memories, called "spacers", in the form of short viral genomic sequences which they store in CRISPR arrays in their own genomes [7–9]. These sequences are then transcribed and processed into short crRNA fragments that guide CRISPR-associated (Cas) proteins to the target viral sequences (or "protospacers") so that the foreign DNA or RNA can be degraded [9–11]. Thus the CRISPR array is the genomic location in which memories are recorded, while the Cas proteins act as the machinery of the immune system, with specific proteins implicated in memory acquisition, crRNA processing, or immune targeting.

CRISPR systems appear to be widespread across diverse bacterial and archaeal lineages, with previous analyses of genomic databases indicating that $\sim 40\%$ of bacteria and $\sim 80\%$ of archaea have at least one CRISPR system [12–14]. These systems vary

widely in *cas* gene content and targeting mechanism, although the *cas1* and *cas2* genes involved in spacer acquisition are universally required for a system to be fully functional [9, 12]. Such prevalence suggests that CRISPR systems effectively defend against phage in a broad array of environments. The complete story seems to be more complicated, with recent analyses of environmental samples revealing that some major bacterial lineages almost completely lack CRISPR systems and that the distribution of CRISPR systems across prokaryotic lineages is highly uneven [15]. Other studies suggest that particular environmental factors can be important in determining whether or not CRISPR immunity is effective (e.g., in thermophilic environments [16, 17]). While previous work has focused on the presence or absence of CRISPR across lineages and habitats, little attention has been paid to the number of systems in a genome.

In fact, the multiplicity of CRISPR systems per individual genome varies greatly, with many bacteria having multiple CRISPR arrays and some having multiple sets of *cas* genes as well (e.g., [18, 19]). CRISPR and other immune systems are horizontally transferred at a high rate relative to other genes in bacteria [20], meaning that any apparent redundancy of systems may simply be the result of the selectively neutral accumulation of systems within a genome. Alternatively, there are a number of reasons, discussed below, why having multiple sets of *cas* genes or CRISPR arrays might be adaptive.

We suspected that possessing multiple CRISPR systems might provide a selective advantage, given that the phenomenon is common across prokaryotes (as detailed below) and, in some clades, appears to be conserved over evolutionary time (e.g. [21, 22]). Since microbial genomes have a deletion bias [23, 24], we would expect extraneous systems to be removed over time. Here we construct a test of neutral CRISPR array accumulation via horizontal transfer and loss. Using publicly available genome data we show that the number of CRISPR arrays in a wide range of prokaryotic lineages deviates from this neutral expectation, including in the case that one array is assumed to be selectively maintained. Thus we conclude that, on average, prokaryotes are under selection to have multiple CRISPR arrays. We go on to discuss several hypotheses for why having multiple arrays might be adaptive. Finally, we suggest that a tradeoff between the rate of acquisition of immune memory and the span of immune memory could lead to selection for multiple CRISPR arrays.

Materials and methods

Dataset

All available completely sequenced prokaryotic genomes (all assembly levels) were downloaded from NCBI's non-redundant RefSeq database FTP site (`ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria`, [25]) on May 11, 2017. Genomes were scanned for the presence of CRISPR arrays using the CRISPRDetect software [26]. We used default settings except that we did not take the presence of *cas* genes into account in the scoring algorithm (to avoid circularity in our arguments), and accordingly used a quality score cutoff of three, following the recommendations in the CRISPRDetect documentation. CRISPRDetect also identifies the consensus repeat sequence and determines the number of repeats for each array. Presence or absence of *cas* genes were determined using genome annotations from NCBI's automated genome annotation pipeline for prokaryotic genomes [27]. We discarded genomes that lacked a CRISPR array in any known members of their taxon. In this way we only examined genomes known to be compatible with CRISPR immunity.

Test for adaptiveness

Our power to detect adaptation hinges on our ability to differentiate between non-functional (i.e., neutrally-evolving) and functional (i.e., potentially-selected) CRISPR arrays. Since all known CRISPR systems require the presence of *cas1* and *cas2* genes in order to acquire new spacers, we use the presence of both genes as a marker for functionality and the absence of one or both genes as a marker for non-functionality. Henceforth we will consider CRISPR arrays in genomes with both *cas1* and *cas2* genes to be "functional" and CRISPR arrays in genomes lacking either *cas1*, *cas2*, or both genes to be "non-functional". This differentiation allows us to consider the probability distributions of the number of CRISPR arrays i in non-functional (N_i) and functional (F_i) genomes, respectively.

We start with our null hypothesis, that in genomes with functional CRISPR systems possession of a single array is highly adaptive (i.e. viruses are present and will kill any susceptible host), but that additional arrays provide no additional advantage. Thus these additional arrays will appear and disappear in a genome as the result of a neutral birth/death horizontal transfer and loss process, where losses are assumed to remove an array in its entirety. This hypothesis predicts that the non-functional distribution will look like the functional distribution shifted by one (S_i):

$$H_0 : N_i \approx S_i = F_{i+1} / \sum_{j=1}^{\infty} F_j \tag{1}$$

for $i \geq 1$. We take two approaches to testing this prediction: one parametric from first principles with greater power but more assumptions and one non-parametric with less power but also fewer assumptions.

We begin by deriving a functional form for the distribution N_i from first principles following a neutral process. If CRISPR arrays arrive in a given genome at a constant rate via rare horizontal transfer events, then we can model their arrivals using a Poisson process with rate η. Assuming arrays are also lost independently at a constant rate, the lifetime of each array in the genome will be exponentially distributed with rate ν. This leads to a linear birth-death process of array accumulation, which yields a Poisson equilibrium distribution with rate $\lambda = \frac{\eta}{\nu}$. While this rate might be constant for a given taxon, it will certainly vary across taxa due to different intrinsic (e.g. cell wall and membrane structure) and extrinsic factors (e.g. density of neighbors, environmental pH and temperature) [20]. We model this variation by allowing genome j to have rate $\lambda_j = \frac{\eta_j}{\nu_j}$ and assuming $\lambda_j \sim \text{Gamma}(\alpha, \beta)$, which we pick for its flexibility and analytic tractability. This combination of gamma and Poisson distributions leads to the number of arrays i in a random genome following a negative binomial distribution $N_i = \text{NB}(r, p)$ where $r = \alpha$ and $p = \frac{\beta}{1+\beta}$.

Now we can fit this distribution to data to find maximum likelihood estimates of r and p for the distribution of array counts in both the set of non-functional genomes (N_i) and the set of functional genomes as shifted under our null hypothesis (S_i). This allows us to construct a parametric test of multi-array adaptiveness. We expect that $\hat{r}_N \approx \hat{r}_S$ and $\hat{p}_N \approx \hat{p}_S$ under our null hypothesis (where subscripts correspond to the distribution to which the parameters were fit). When our null hypothesis is violated it should shift the means of these distributions. Therefore we estimate and compare these means $\mu_k = \frac{p_k r_k}{1-p_k}$, $k \in N, S$. We expect that $\hat{\mu}_S > \hat{\mu}_N$ if more than one array is adaptive, and we bootstrap confidence intervals on these estimates to determine whether the effect is significant. A similar non-parametric version of this test is to simply compare sample means, but at a sacrifice of power.

We also construct a non-parametric test of adaptiveness by determining at what shift s the mismatch between $F_{i+s} / \sum_{j=1+s}^{\infty} F_j$ and N_i, measured as the sum of squared

differences between the distributions, is minimized:

$$s^\star = \underset{s}{\operatorname{argmin}} \sum_{i=0}^{\infty} \left(N_i - F_{i+s} / \sum_{j=s}^{\infty} F_j \right)^2 . \tag{2}$$

Under our null hypothesis $s^\star = 1$, and a value of $s^\star > 1$ implies that having more than one array is adaptive. Our parametric test is superior to $s^\star$ because it can detect if having more than one array is adaptive across the population on average, but not in all taxa, so that the optimal shift is fractional.

Correcting for correlations in HGT

Differential rates of horizontal gene transfer (HGT) between lineages could produce an observed correlation between *cas1* and *cas2* presence and array count in the absence of any selection for having multiple CRISPR arrays. In other words, some genomes would be functional and have many arrays due to a high arrival rate of foreign genetic material, and other lineages would be non-functional and lack CRISPR arrays simply because of low rates of HGT. If this were the case, then comparisons between these lineages would lead to a spurious result of adaptiveness.

There are several ways to control for such correlation. First, if HGT differences among lineages can explain any *cas*-CRISPR correlation, then beyond simple presence or absence of *cas* genes we should see that an increased number of *cas* genes in a genome is associated with an increased number of arrays. We can differentiate between the two by plotting the number of *cas1* genes in a genome against the number of arrays, excluding those genomes lacking *cas1* to control for the potential effects of CRISPR adaptiveness on *cas1* presence/absence. Second, we can perform our parameter-based test on a subset of the data such that we take an equal number of functional and non-functional genomes from each species to control for lineage-specific effects. Finally, we can also perform a species-wise parameter-based test. In this case, for each species k we calculate $\Delta\mu_k = \hat{\mu}_{S_k} - \hat{\mu}_{N_k}$ and then bootstrap the mean of the distribution of these values ($\bar{\Delta\mu_k}$) to detect if there is a significant difference from zero.

CRISPR spacer turnover model

We develop a simple deterministic model of the spacer turnover dynamics in a single CRISPR array of a bacterium exposed to n viral species (i.e., disjoint protospacer sets):

$$\underbrace{\frac{dC_i}{dt}}_{\text{Spacers Targeting Viral Sp. } i} = \underbrace{a_i(t, C_i)}_{\text{Acquisition}} - \underbrace{\mu_L C_i \sum_j C_j}_{\text{Loss}} \tag{3}$$

where μ_L is the spacer loss rate parameter and a_i is a function of time representing the viral environment. Here we let $a_i(t, C_i) = \mu_A v_i f_i(t)$, where μ_A is the spacer acquisition rate, v_i is a composite parameter describing the maximum density of a viral species in the environment multiplied by the adsorption rate, and $f_i(t)$ is a function describing the fluctuations of the viral population over time that takes values between zero and one.

The rate of per-spacer loss increases linearly with locus length. This assumption is based on the observation that spacer loss appears to occur via homologous recombination between repeats [28–30], which becomes more likely with increasing numbers of spacers (and thus repeats). Using this model we can determine optimal spacer acquisition rates given a particular pathogenic environment. If there are multiple optima, or if optima cluster in different regions of parameter space for different

pathogenic environments, this indicates that having multiple-arrays may be the best solution in a given environment or set of environments.

We analyze a simple case with two viral species where there is one "background" species (B) representing the set of all viruses persisting over time in the environment ($f_B(t) = 1$) and another "transient" species (T) that leaves and returns to the environment after some interval of time ($f_T(t)$ is a binary function that takes a value of one if virus T is present in the environment and zero otherwise). This allows us to effectively illustrate any tradeoff between the ability to maintain defenses towards a constant threat and the ability to defend against threats that periodically reappear in the environment. In practice we will focus on one interval in which T leaves and then returns to the system in order to see if immune memory is lost, and if so how long it takes to regain that memory, given a particular spacer acquisition rate (μ_A). We also consider how long it would take to acquire immunity towards a completely novel phage species given a particular acquisition rate in order to assess any tradeoff between learning-speed and memory-span.

We also can consider the phenomenon of priming in our model, wherein if a CRISPR array has a spacer targeting a particular viral species, the rate of spacer acquisition towards that species is increased [31, 32]. Thus

$$a_i(t, C_i) = \mu_A v_i f_i(t) g(C_i) \tag{4}$$

where

$$g(C_i) = \begin{cases} 1 & C_i < 1 \\ p & C_i \geq 1 \end{cases} \tag{5}$$

is a stepwise function determining the presence or absence of at least one spacer towards a given viral species and $p > 1$ is the degree of priming. For details of model analysis see S1 Text.

There is evidence that mature crRNA transcripts from the leading end of the CRISPR array are far more abundant than those from the trailing end, and that this decay over the array happens quickly (most transcripts from within the first few spacers, though some spacers farther down can occasionally show high representation) [33–35]. This suggests an alternative model wherein the length of an "effective" array is capped at a constant number of spacers and "loss" (i.e. movement out of the zone where crRNAs are actively transcribed and processed) occurs directly due to the acquisition of novel spacers. Therefore we built a modified version of the model above with a hard cap on array length (S1 Text).

Results

Having more than one CRISPR array is common

About half of the prokaryotic genomes in the RefSeq database have at least one CRISPR array (44%). Of these genomes, almost half have more than one CRISPR array (48%). When restricting ourselves only to putatively functional genomes where the CRISPR spacer acquisition machinery was present (*cas1* and *cas2*) the proportion of genomes with more than one array increases to 64%. In contrast to this result, having more than one set of *cas* targeting genes is not nearly as common. Signature targeting genes are diagnostic of CRISPR system type. We counted the number of signature targeting genes for type I, II, and III systems in each genome that had at least one CRISPR array (*cas3*, *cas9*, and *cas10* respectively [36]). Only 2% of these genomes have more than one targeting gene (either multiple copies of a single type or multiple types). Even when restricting ourselves again to genomes with intact acquisition

machinery, only 3% of genomes had multiple signature targeting genes. However, of those genomes with more than one set of *cas* genes, most had multiple types (80%).

Some taxa are overrepresented in RefSeq (e.g. because of medical relevance), and we wanted to avoid results being driven by just those few particular taxa. We controlled for this by randomly sub-sampling 10 genomes from each taxa with greater than 10 genomes in the database. After sub-sampling, approximately 41% of genomes had at least one CRISPR array, and of these 47% had more than one. Of genomes with intact spacer acquisition machinery, 62% had more than one CRISPR array. A larger fraction of these sub-sampled genomes had more than one set of *cas* targeting genes when at least one CRISPR array was present (9%), indicating that most highly-represented species did not possess multiple sets of cas targeting genes. Of these multi-*cas* genomes, most had multiple types (84%).

Validation of functional / non-functional classification

Our power to detect selection depends critically on our ability to classify genomes as CRISPR functional vs. non-functional. Functional CRISPR arrays should, on average, contain more spacers than non-functional arrays. Thus we compared the number of repeats in CRISPR arrays in genomes with both *cas1* and *cas2* present ("functional") to the number of spacers in genomes lacking either or both genes ("non-functional") and confirmed that the former has significantly more than the latter ($t = -31.29$, $df = 42562$, $p < 2.2 \times 10^{-16}$) (S1 Fig, [37]). This difference in length (3.88) is not as large as one might expect, possibly because some systems are able to acquire or duplicate spacers via homologous recombination [38] and arrays may have been inherited recently from strains with active *cas* machinery.

Instantaneous array loss vs. gradual decay

There are two possible routes to complete CRISPR array loss: (1) an all-at-once loss of the array (e.g. due to recombination between flanking insertion sequences [39, 40]) and (2) gradual decay due to spacer loss via homologous recombination [28–30]. Limited experimental evidence supports (1) spontaneous loss of the entire CRISPR array [41], as do comparisons between closely related genomes [40]. The distinction above is important, because if CRISPR array loss were to occur primarily via (2) gradual decay, then non-functional genomes would have an intrinsically higher rate of array loss than functional genomes. This could lead us to spuriously accept a result of multi-array adaptiveness. If arrays were primarily lost via gradual decay we would expect a positive relationship between the number of arrays in a genome and the average array length in a genome. In functional genomes with the complete spacer acquisition machinery (*cas1* and *cas2*) this trend would be due to the higher probability of stochastically reaching a 0-spacers state in shorter arrays, and arrays will in general be shorter in genomes with lower spacer acquisition rates. In non-functional genomes which lack the complete spacer acquisition machinery this trend would result from differences in time since loss of acquisition machinery, where genomes that had lost that machinery farther in the past would have both shorter arrays and fewer arrays on average. Surprisingly, in both cases we find a slightly negative linear relationship between mean array length in a genome and the number of arrays in a genome (functional genomes: $R^2 = 0.02173$, $p < 2 \times 10^{-16}$; non-functional genomes: $R^2 = 0.0002704$, $p = 0.01697$). Thus we reject any positive array-length vs.array count relationship and accordingly rule out array loss via spacer-wise decay.

Having more than one CRISPR array is adaptive

We leveraged the difference between functional and non-functional genomes, each in which the process of CRISPR array accumulation should be distinct (Fig. 1, Table 1). Non-functional CRISPR arrays should accumulate neutrally in a genome following background rates of horizontal gene transfer and gene loss. We constructed two point estimates of this background accumulation process using our parametric model to infer the distribution of the number of arrays. One estimate came directly from the non-functional genomes ($\hat{\mu}_N$, Fig. 1a). The other came from the functional genomes, assuming that having one array is adaptive in these genomes, but that additional arrays accumulate neutrally ($\hat{\mu}_S$, Fig. 1b). If having multiple (functional) arrays is adaptive, then we should find that $\hat{\mu}_N < \hat{\mu}_S$. We found this to be overwhelmingly true, with about two arrays on average seeming to be evolutionarily maintained across prokaryotic taxa ($\Delta\mu = \hat{\mu}_S - \hat{\mu}_N = 1.01 \pm 0.03$, $s^\star = 2$). We bootstrapped 95% confidence intervals of our estimates (Table 1) and found that the bootstrapped distributions did not overlap, indicating a highly significant result (Fig. 1d)

Sub-sampling overrepresented taxa altered our parameter estimates slightly, but did not change our overall result ($\Delta\mu = 0.99 \pm 0.09$, S2 Fig). To control for the possibility that multiple sets of *cas* genes in a small subset of genomes could be driving this adaptive signature, we restricted our dataset only to genomes with one or fewer signature targeting genes (*cas3*, *cas9*, or *cas10* [12, 36]) and one or fewer copies each of the genes necessary for spacer acquisition (*cas1* and *cas2*). Even when restricting our analyses to genomes with one or fewer sets of *cas* genes, it is clearly adaptive to have more than one (functional) CRISPR array, though the effect size is smaller ($\Delta\mu = 0.89 \pm 0.03$, S3 Fig; with sub-sampling of overrepresented taxa $\Delta\mu = 0.57 \pm 0.09$, S4 Fig).

To control for the possibly confounding effects of differences in the rate of HGT between lineages, we performed three additional analyses. First, beyond the clear effect of the presence of at least one set of *cas* genes on the number of arrays in a genome, additional copies of *cas1* genes do not have any strong effect on the number of arrays in a genome (S5 Fig). Second, if we further restrict our sub-sampled dataset to genomes with one or fewer sets of *cas* genes, such that each species is represented by an equal number of functional and non-functional genomes, then we still find a positive signature of adaptiveness ($\Delta\mu = 0.53 \pm 0.16$, S6 Fig). Unfortunately this method involves excluding a large portion of the dataset. Third, our species-wise implementation of the $\Delta\mu$ test that controls for differences in rates of HGT between lineages also confirms a signature of multi-array adaptiveness, though the effect is less strong ($\overline{\Delta\mu}_k = 0.44 \pm 0.14$). Because there is a low number of genomes for most species and this test restricts us to only within-species comparisons, our species-wise parameter-based test lacks power.

Finally, to confirm that assembly level had no effect on our outcome, we ran our parametric test on the set of completely assembled genomes in the dataset. This reduced dataset gave a signature of multi-array adaptiveness, albeit to a lesser degree than the full dataset (4468 genomes, $\Delta\mu = 0.43 \pm 0.07$, $s^* = 2$)

Evidence for array specialization

In genomes with multiple arrays, the dissimilarity between consensus repeat sequences of arrays in a single genome spanned a wide range of values (Levenshtein Distance, S7 Fig and S8 Fig), though the mode was at zero (i.e., identical consensus repeats). When limiting our scope to only genomes with exactly two CRISPR arrays, we saw a bimodal distribution of consensus repeat dissimilarity, with one peak corresponding to identical arrays within a genome and the other corresponding to arrays with essentially randomly

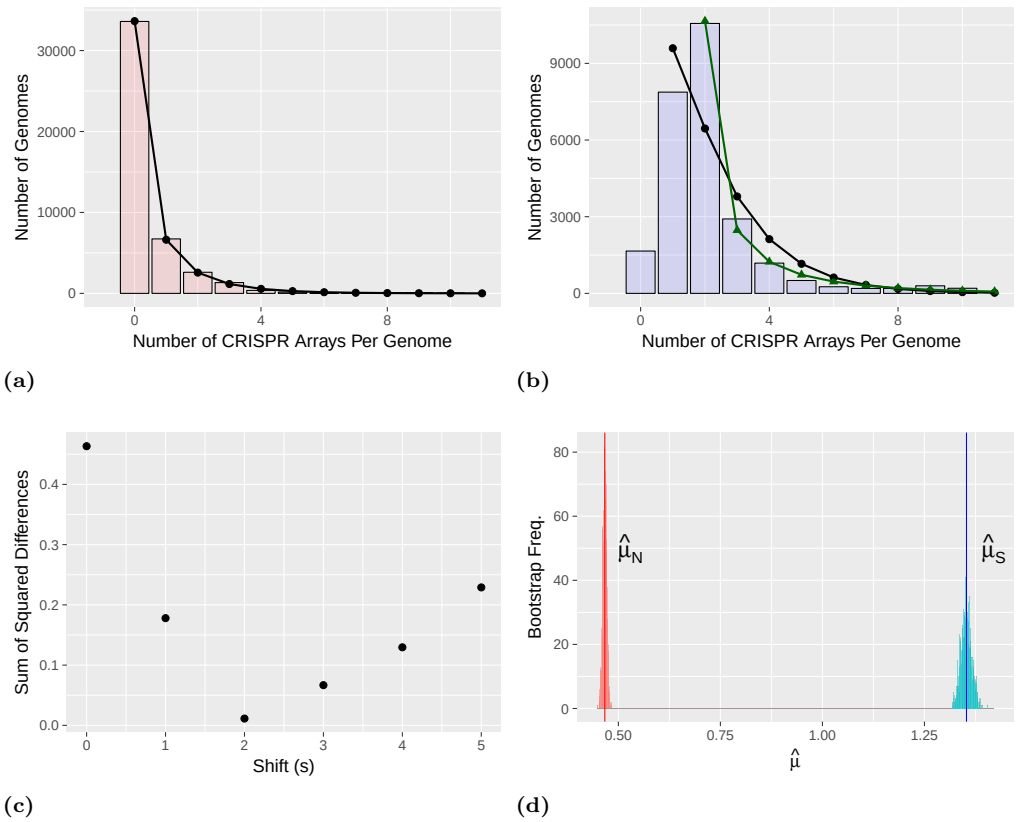


Fig 1. Having more than one CRISPR array is adaptive on average across prokaryotes. (a-b) Distribution of number of arrays per genome in (a) genomes with non-functional CRISPR immunity and (b) genomes with putatively functional CRISPR immunity. The tails of these distributions are cut off for ease of visual comparison (33 genomes with > 10 arrays in (a) and 340 genomes with > 10 arrays in (b)). In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and green triangles to the double-shifted distribution ($s = 2$). (c) The optimal shift is $s^\star = 2$, where the difference between the two distributions is minimized. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show no overlap with 1000 bootstrap replicates.

			Bootstrap			Bootstrap			
Only ≤ 1 *cas* set	Sub-sampled	$\hat{\mu}_S$	2.5%	97.5%	$\hat{\mu}_N$	2.5%	97.5%	$\Delta\mu$	$s^\star$
No	No	1.41	1.45	1.51	0.47	0.46	0.48	1.01	2
No	Yes	2.2	2.12	2.28	1.21	1.15	1.26	0.99	2
Yes	No	1.35	1.33	1.38	0.47	0.46	0.48	0.89	2
Yes	Yes	1.75	1.67	1.82	1.18	1.13	1.23	0.57	2

Table 1. Tests for multi-array adaptiveness applied to different subsets of the RefSeq data. See Fig 1 and S2 Fig-S4 Fig.

drawn repeat sequences except for a few conserved sites between them (S7D Fig). We also observed that among functional genomes, the area of the peak corresponding to dissimilar repeat sequences was significantly higher than among non-functional genomes ($\chi^2 = 16.784$, $df = 1$, $p < 4.19 \times 10^{-5}$, S7 Fig). This suggests that the observed signature of adaptiveness may be related to the diversity of consensus repeat sequences among CRISPR arrays in a genome.

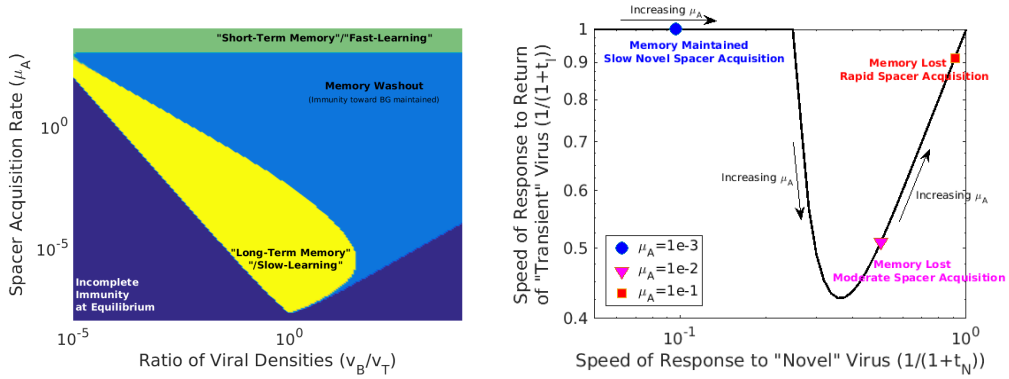
We next sought to assess if this observed variability in repeat sequences among arrays might have functional implications for CRISPR immunity, even when arrays share a set of *cas* genes. One measure of system functionality is array length, as we expect it to be correlated with the rate of spacer acquisition. Therefore, we determined whether the mean pairwise dissimilarity between array consensus repeat sequences in a genome was associated with the variance of array lengths in that genome. Array length was measured as the number of repeats in an array. Again we used our dataset restricted to genomes with one set of *cas* genes and with sub-sampled genomes. The mean pairwise distance between consensus repeats within a genome was positively associated with the variance of the number of repeats across arrays in a genome, but this relationship was poorly predictive and likely spurious ($R^2 = 0.007464$, $p < 0.00123$).

A tradeoff between memory span and acquisition rate could select for multiple arrays in a genome

The evidence above shows that functional genomes have more diverse repeat sequences among their arrays than non-functional genomes do. We hypothesized that having multiple systems with different acquisition rates could allow prokaryotes to respond to a range of pathogens with different characteristics (e.g. residence time in the environment, frequency of recurrence). To investigate this possibility we built a simple model of spacer turnover dynamics in a single CRISPR array in the presence of "background" and "transient" viral species (see Methods). We constructed phase diagrams of the model behavior, varying spacer acquisition rates and the relative population sizes of viral species or the extent of priming, respectively (Fig. 2a, S9 Fig). We found that for very high spacer acquisition rates, the system is able to maintain immunity to both background and transient viral populations ("short-term memory"/"fast-learning"). High rates of spacer acquisition are unrealistic as they lead to high rates of autoimmunity ([42–46], S2 Text). Our analysis also reveals that there is a region of parameter space with low spacer acquisition rates in which immunity is maintained ("long-term memory"/"slow-learning"). This is the region where low spacer turnover rates allow immune memory to remain in the system over longer periods of time (Fig. 2a).

The "long-term memory"/"slow-learning" region of parameter space is separated from the "short-term memory"/"fast-learning" region of parameter space by a "memory-washout" region in which spacer turnover is high so that memory is lost but acquisition is not rapid enough to quickly re-acquire immunity towards the transient virus (Fig. 2a). The relative densities of the different viral species modulate the relative importance of fast-acquisition versus memory span. Thus for a range of pathogenic environments two distinct CRISPR immune strategies exist with respect to the spacer acquisition rate ("long-term memory" vs. "fast-learning"). We also note that high levels of priming expand the "washout" region separating the two strategies, as high spacer uptake from background viruses will crowd out long term immune memory (S9 Fig).

A single CRISPR array fulfilling one of the two CRISPR strategies is sufficient in the case shown in Fig 2a, although only one of those strategies is likely accessible due to limits on spacer acquisition rates. When a third, novel viral species enters the system, a two-array solution may become necessary due to the memory span vs. learning speed



(a) **(b)**

Fig 2. Immune memory is maximized at intermediate and low spacer acquisition rates, creating a tradeoff with the speed of immune response to novel threats. (a) Phase diagram of the behavior of our CRISPR array model with two viral species, a constant "background" population and a "transient" population that leaves and returns to the system at some fixed interval. The yellow region indicates that immunity towards both viral species was maintained. The green region indicates where immune memory was lost towards the transient phage species, but reacquired almost immediately upon phage reintroduction ($t_I < 10^{-5}$, where t_I is the time to first spacer acquisition after the return of the species to the system following an interval of absence). The light blue region indicates that only immunity towards the background species was maintained (i.e., immune memory was rapidly lost and $t_I > 10^{-5}$). Dark blue indicates where equilibrium spacer content towards one or both species did not exceed one despite both species being present in the system (S1 Text). (b) The tradeoff between memory span and learning speed. The speed of immune response to the transient viral species (as in (a), with $v_B/v_F = 1$) is plotted against the speed of response to a novel viral species to which the system has not been previously exposed (so that there are no spacers targeting this species), over a range of μ_A values ($\mu_A \in [10^{-3.5}, 1]$). The speed of immune response to the transient species is defined as $1/(1 + t_I)$ (where $t_I = 0$ if memory is maintained). The speed of response to the novel species is similarly defined as $1/(1 + t_N)$ where t_N is the time to first spacer acquisition towards this species. For specifics on calculating t_I and t_N see S1 Text.

tradeoff (Fig 2b). Extremely high spacer acquisition rates might allow the host to rapidly respond to both novel and returning threats, but, as noted above, such rates are unrealistic due to physical constraints on the speed of adaptation as well as the evolutionary constraint of autoimmunity ([42–46], S2 Text). CRISPR adaptation is rapid, but it is not instantaneous, and infected hosts with arrays lacking an appropriate spacer will often perish before a spacer can be acquired [47]. This means that the region of rapid immune response to both transient and novel threats by a single array, shown in the top-right corner of Fig 2b, is probably inaccessible. Thus, in order to maximize novel spacer acquisition and memory span simultaneously, a two-system solution will be required.

An alternative model with fixed-length arrays, corresponding to a situation in which only the first few spacers in an array are processed into mature crRNAs [33–35], demonstrated qualitatively similar behavior to the model described above (S1 Text, S10 Fig).

Taxon-specific signatures of adaptiveness

Several taxa in the dataset were represented by a sufficiently large number of genomes (> 1000) that varied in the presence of both *cas* genes and CRISPR array counts that we were able to reliably perform our test for adaptiveness on each of these taxa individually. We found that among *Klebsiella pneumoniae* and *Staphylococcus aureus* genomes there was a signal of multi-system adaptiveness ($\Delta\mu = 0.60 \pm 0.06$, 0.63 ± 0.20 respectively), though relatively few of the *S. aureus* genomes were functional (0.5%). *Pseudomonas aeruginosa* showed no signal of multi-array adaptiveness ($\Delta\mu = 0.15 \pm 0.17$), and *Escherichia coli* and *Mycobacterium tuberculosis* both showed very weak signals ($\Delta\mu = 0.09 \pm 0.06$, 0.12 ± 0.05 respectively), indicating that these species may occupy niches that favor single-array strains. *Salmonella enterica* had strongly negative $\Delta\mu$ values ($\Delta\mu = -1.05 \pm 0.11$), indicating that functional arrays are selected against in this taxon. Previous work has shown that CRISPR in *E. coli* and *S. enterica* appears to be non-functional as an immune system under natural conditions [48, 49]. All of these taxa are human pathogens, and can occupy a diverse set of environmental niches on the human body. It is unclear at this time what is causing the differences in the adaptive landscape each taxon experiences.

A very small portion of the genomes used in our analyses were from archaea ($< 1\%$). We ran our analyses on these genomes alone to see if they differed significantly from their bacterial counterparts. No signature of multi-array adaptiveness was detected, although we note that the large majority of genomes had CRISPR arrays and were also functional, making our approach less powerful (S11 Fig). Further, if those few non-functional genomes lost their *cas* spacer acquisition machinery recently, then our power would be reduced even more because these genomes might still bear the remnants of past selection.

Discussion

Having multiple CRISPR arrays is adaptive across prokaryotic taxa

On average, having more than one CRISPR array is adaptive. This surprising result holds controlling for both overrepresented taxa and the influence of multiple sets of *cas* genes. However, the degree of adaptation appears to vary between taxa, likely as a function of the pathogenic environment each experiences based on its ecological niche. Additionally, we showed that arrays in functional genomes are more diverse than in

those non-functional genomes, indicating that array diversity may be important in addition to array multiplicity.

The data appear to follow a negative binomial distribution quite well (Figs 1b and 1a, S2 Fig-S4 Fig), consistent with our theoretical prediction. This pattern is robust to sub-setting of the data in a variety of ways. We note that, due to the large size of this dataset, formal goodness-of-fit tests to the negative binomial distribution always reject the fit due to small but statistically significant divergences from the theoretical expectation.

Our test for adaptiveness is conservative to the miscategorization of arrays as "functional"or "non-functional". Miscategorizations could occur because intact targeting machinery still allows for preexisting spacers to confer immunity, some CRISPR arrays may be conserved for non-immune purposes (e.g. [48, 50]), or intact acquisition machinery is no guarantee of system functionality. That being said, our test is conservative precisely because of such miscategorizations, as they should increase $\hat{\mu}_N$ and decrease $\hat{\mu}_S$ respectively. Selection against having a CRISPR array in genomes lacking spacer acquisition machinery could produce a false positive signature of adaptiveness. This is unlikely because there is no reason a non-functional CRISPR array should be under strong negative selection given the low or nonexistent associated costs.

Finally, we note that μ_N and μ_S take on a range of values depending on what subset of taxa/genomes are included in the dataset. This is to be expected as each set of species will occupy a distinct environment in terms of both the rate of horizontal gene transfer and the usefulness of CRISPR immunity. Nevertheless, our qualitative result of multi-CRISPR adaptiveness is robust to this quantitative variability.

Why have two CRISPR-Cas systems?

Our data show significant numbers of both similar and dissimilar CRISPR arrays within the same genome, so either could potentially be adaptive. While CRISPR systems are generally highly flexible, a prokaryote might still gain an advantage in the former case if multiple similar systems lead to improved immunity through redundancy and in the latter case if multiple dissimilar systems allow for specialization towards multiple types of threats. The relevance of the different advantages depends on whether an individual has multiple sets of *cas* genes, CRISPR arrays, or both.

In the case of similar systems, immunity could be improved by (a) an increased spacer acquisition rate, (b) an increased rate of targeting, or (c) a longer time to expected loss of immunity. Duplication of *cas* genes could, in principle, increase uptake (a) and targeting rates (b) through increased gene expression, but our data show that multiple sets of *cas* genes are rare, which suggests this is, at best, a minor force. Alternatively, duplication of CRISPR arrays could increase targeting (b) via an increased number of crRNA transcripts or increase memory duration (c) through spacer redundancy. However, the effectiveness of crRNA may actually decrease in the presence of competing crRNAs [51, 52] and, since a single array can have multiple spacers with the same target, there is a diminished advantage to having multiple arrays in terms of memory span (S3 Text). Redundant arrays might also be a form of bet-hedging since CRISPR functionality is lost at a high rate in some prokaryotes [41]. While this last explanation is plausible, it does not explain why functional genomes would have more diverse arrays than non-functional genomes.

In the case of dissimilar systems, immunity could be aided if diverse features are advantageous. For example some viruses encode proteins that deactivate Cas targeting proteins [53–55]. Diverse *cas* genes may allow hosts to evade the action of these anti-CRISPR proteins, which are often extremely broadly acting [53, 54]. Alternatively, it has recently been shown that promiscuous type III Cas proteins are often encoded alongside type I systems and can function as a "backup", using spacers from the same

array to target phages that have mutated the protospacer-adjacent motifs necessary for type I targeting [56]. Most genomes with multiple *cas* signature genes also had multiple types of such genes, suggesting some diversifying force. The inclusion of these multi-*cas* genomes also increased the effect size of our test for adaptiveness, despite low representation in the dataset. In any case, while these hypotheses remain interesting candidates to explain CRISPR multiplicity in some prokaryotes, the majority of the genomes in the dataset have only one set of *cas* genes and thus these mechanisms cannot explain the signature for multi-array adaptiveness observed in the majority of the dataset.

Our models suggest that dissimilarity between systems can also lead to diverse spacer acquisition rates. We show that a tradeoff between memory span and learning speed leads to selection for both high acquisition rate (i.e., short term memory) and low acquisition rate (i.e., long-term memory) systems, depending on the pathogenic environment of the host. As an array increases in length (i.e., the number of repeats increases) the rate of spacer loss should also increase because loss occurs via homologous recombination. A length-dependent spacer loss rate causes high acquisition rate systems to also have high loss rates, producing the aforementioned tradeoff. Even CRISPR arrays sharing a single set of *cas* genes may vary greatly in acquisition rate [57]. Similarly, arrays with slightly different consensus repeat sequences may differ in length, despite sharing a set of *cas* genes [58], which suggests a functional role for repeat sequence modifications in determining spacer insertion rates. Despite this, we found no clear link between array length and repeat sequence in our data, possibly due to a weakly array-length dependent spacer loss rate. A slightly modified model that imposes a hard cap on the number of "effective" spacers in an array (S1 Text) and would not require a strong array-length vs. acquisition rate link produced qualitatively similar results. We based this model on evidence that the abundance of mature crRNA transcripts from the leader end greatly exceeds that of those from the trailing end [33–35], implying that only the first few spacers in an array will provide effective immunity. We speculate that if Cas acquisition and insertion proteins are flexible to some degree in the repeat sequences they recognize, then certain sequences may be favored over others, producing the variability in acquisition rate required for the tradeoff we propose to occur.

Experimental verification that the consensus repeat sequence modulates spacer acquisition rates will be a first step towards validating our proposed tradeoff mechanism. As more genome sequences from environmental samples become available, it will be possible to explicitly link particular array configurations to specific features of the pathogenic environment or host lifestyle. Even then, open questions remain. One phenomenon that we do not address here is that a small, but non-trivial number of genomes have greater than 10 arrays. It is difficult to imagine so many arrays accumulating neutrally in a genome. If high array counts are a product of high horizontal transfer rates, then genomes with extremely high array counts should also be larger due to accumulation of foreign genetic material. This was not the case (S12 Fig), indicating that rates of horizontal transfer alone cannot explain these outliers.

Finally, our examination of immune configuration is likely relevant to the full range of prokaryotic defense mechanisms. In contrast to previous work focusing on mechanistic diversity (e.g. [4, 16, 43, 59]), we emphasize the importance of the multiplicity of immune systems in the evolution of host defense. As we suggest, a surprising amount of strategic diversity may masquerade as simple redundancy.

Acknowledgments

JLW was supported by a GAANN Fellowship from the U.S. Department of Education and the University of Maryland. WFF was partially supported the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-14-1-0490. PLFJ was supported in part by NIH R00 GM104158.

References

1. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. Journal of Bacteriology. 2011;193(21):6039–6056. doi:10.1128/JB.05535-11.

2. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Research. 2013;41(8):4360–4377. doi:10.1093/nar/gkt157.

3. Houte Sv, Buckling A, Westra ER. Evolutionary Ecology of Prokaryotic Immune Mechanisms. Microbiology and Molecular Biology Reviews. 2016;80(3):745–763. doi:10.1128/MMBR.00011-16.

4. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. BMC Evolutionary Biology. 2015;15:43. doi:10.1186/s12862-015-0324-2.

5. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology Direct. 2006;1:7. doi:10.1186/1745-6150-1-7.

6. Goren M, Yosef I, Edgar R, Qimron U. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. RNA biology. 2012;9(5):549–554. doi:10.4161/rna.20177.

7. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution. 2005;60(2):174–182. doi:10.1007/s00239-004-0046-3.

8. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005;151(8):2551–2561. doi:10.1099/mic.0.28048-0.

9. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007;315(5819):1709–1712. doi:10.1126/science.1138140.

10. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science (New York, NY). 2008;322(5909):1843–1845. doi:10.1126/science.1165771.

11. Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015;526(7571):55–61. doi:10.1038/nature15386.

12. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology. 2011;9(6):467–477. doi:10.1038/nrmicro2577.

13. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry. 2013;82(1):237–266. doi:10.1146/annurev-biochem-072911-172315.

14. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, et al. New CRISPR–Cas systems from uncultivated microbes. Nature. 2017;542(7640):237–241. doi:10.1038/nature21059.

15. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications. 2016;7:10613. doi:10.1038/ncomms10613.

16. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary Dynamics of the Prokaryotic Adaptive Immunity System CRISPR-Cas in an Explicit Ecological Context. Journal of Bacteriology. 2013;195(17):3834–3844. doi:10.1128/JB.00412-13.

17. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV. Viral Diversity Threshold for Adaptive Immunity in Prokaryotes. mBio. 2012;3(6):e00456–12. doi:10.1128/mBio.00456-12.

18. Horvath P, Coûté-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, Barrangou R. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. International Journal of Food Microbiology. 2009;131(1):62–70. doi:10.1016/j.ijfoodmicro.2008.05.030.

19. Cai F, Axen SD, Kerfeld CA. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. RNA Biology. 2013;10(5):687–693. doi:10.4161/rna.24571.

20. Puigbò P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV. Reconstruction of the evolution of microbial defense systems. BMC Evolutionary Biology. 2017;17:94. doi:10.1186/s12862-017-0942-y.

21. Boudry P, Semenova E, Monot M, Datsenko KA, Lopatina A, Sekulovic O, et al. Function of the CRISPR-Cas System of the Human Pathogen Clostridium difficile. mBio. 2015;6(5):e01112–15. doi:10.1128/mBio.01112-15.

22. Andersen JM, Shoup M, Robinson C, Britton R, Olsen KEP, Barrangou R. CRISPR Diversity and Microevolution in Clostridium difficile. Genome Biology and Evolution. 2016;8(9):2841–2855. doi:10.1093/gbe/evw203.

23. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends in genetics: TIG. 2001;17(10):589–596.

24. Kuo CH, Ochman H. Deletional Bias across the Three Domains of Life. Genome Biology and Evolution. 2009;1:145–152. doi:10.1093/gbe/evp016.

25. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;44(Database issue):D733–D745. doi:10.1093/nar/gkv1189.

26. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17:356. doi:10.1186/s12864-016-2627-0.

27. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research. 2016;44(14):6614–6624. doi:10.1093/nar/gkw569.

28. Garrett RA, Shah SA, Vestergaard G, Deng L, Gudbergsdottir S, Kenchappa CS, et al. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. Biochemical Society Transactions. 2011;39(1):51–57. doi:10.1042/BST0390051.

29. Gudbergsdottir S, Deng L, Chen Z, Jensen JVK, Jensen LR, She Q, et al. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Molecular Microbiology. 2011;79(1):35–49. doi:10.1111/j.1365-2958.2010.07452.x.

30. Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, et al. Persisting Viral Sequences Shape Microbial CRISPR-based Immunity. PLoS Comput Biol. 2012;8(4):e1002475. doi:10.1371/journal.pcbi.1002475.

31. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature Communications. 2012;3:945. doi:10.1038/ncomms1937.

32. Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. CRISPR Interference Directs Strand Specific Spacer Acquisition. PLoS ONE. 2012;7(4):e35888. doi:10.1371/journal.pone.0035888.

33. Bernick DL, Cox CL, Dennis PP, Lowe TM. Comparative genomic and transcriptional analyses of CRISPR systems across the genus Pyrobaculum. Frontiers in Microbiology. 2012;3. doi:10.3389/fmicb.2012.00251.

34. Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, et al. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Molecular Cell. 2012;45(3):292–302. doi:10.1016/j.molcel.2011.10.023.

35. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. Nucleic Acids Research. 2012;40(19):9887–9896. doi:10.1093/nar/gks737.

36. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nature Reviews Microbiology. 2015;13(11):722–736. doi:10.1038/nrmicro3569.

37. Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. The ISME Journal. 2015;9(9):2021–2027. doi:10.1038/ismej.2015.20.

38. Kupczok A, Landan G, Dagan T. The Contribution of Genetic Recombination to CRISPR Array Evolution. Genome Biology and Evolution. 2015;7(7):1925–1939. doi:10.1093/gbe/evv113.

39. Almendros C, Mojica FJ, Díez-Villaseñor C, Guzmán NM, García-Martínez J. CRISPR-Cas functional module exchange in Escherichia coli. MBio. 2014;5(1):e00767–13.

40. Shah SA, Garrett RA. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. Research in Microbiology. 2011;162(1):27–38. doi:10.1016/j.resmic.2010.09.001.

41. Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA. Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. PLoS Genet. 2013;9(9):e1003844. doi:10.1371/journal.pgen.1003844.

42. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. Genes & Development. 2015;29(4):356–361. doi:10.1101/gad.257550.114.

43. Kumar MS, Plotkin JB, Hannenhalli S. Regulated CRISPR Modules Exploit a Dual Defense Strategy of Restriction and Abortive Infection in a Model of Prokaryote-Phage Coevolution. PLoS computational biology. 2015;11(11):e1004603. doi:10.1371/journal.pcbi.1004603.

44. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research. 2012; p. gks216. doi:10.1093/nar/gks216.

45. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015;520(7548):505–510. doi:10.1038/nature14302.

46. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics. 2010;26(8):335–340. doi:10.1016/j.tig.2010.05.008.

47. Hynes AP, Villion M, Moineau S. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nature Communications. 2014;5:4399. doi:10.1038/ncomms5399.

48. Touchon M, Rocha EPC. The Small, Slow and Specialized CRISPR and Anti-CRISPR of Escherichia and Salmonella. PLoS ONE. 2010;5(6):e11126. doi:10.1371/journal.pone.0011126.

49. Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C. CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. Journal of Bacteriology. 2011;193(10):2460–2467. doi:10.1128/JB.01307-10.

50. Li R, Fang L, Tan S, Yu M, Li X, He S, et al. Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. Cell Research. 2016;26(12):1273–1287. doi:10.1038/cr.2016.135.

51. Stachler AE, Marchfelder A. Gene Repression in Haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats) - Cas I-B system. Journal of Biological Chemistry. 2016; p. jbc.M116.724062. doi:10.1074/jbc.M116.724062.

52. Stachler AE, Turgeman-Grott I, Shtifman-Segal E, Allers T, Marchfelder A, Gophna U. High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. Nucleic Acids Research. 2017;doi:10.1093/nar/gkx150.

53. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 2013;493(7432):429–432. doi:10.1038/nature11723.

54. Pawluk A, Bondy-Denomy J, Cheung VHW, Maxwell KL, Davidson AR. A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. mBio. 2014;5(2):e00896–14. doi:10.1128/mBio.00896-14.

55. Rauch BJ, Silvis MR, Hultquist JF, Waters CS, McGregor MJ, Krogan NJ, et al. Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. Cell. 2017;168(1–2):150–158.e10. doi:10.1016/j.cell.2016.12.009.

56. Silas S, Lucas-Elio P, Jackson SA, Aroca-Crevillén A, Hansen LL, Fineran PC, et al. Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. eLife. 2017;6. doi:10.7554/eLife.27601.

57. Staals RHJ, Jackson SA, Biswas A, Brouns SJJ, Brown CM, Fineran PC. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nature Communications. 2016;7:12853. doi:10.1038/ncomms12853.

58. Zeng H, Zhang J, Li C, Xie T, Ling N, Wu Q, et al. The driving force of prophages and CRISPR-Cas system in the evolution of Cronobacter sakazakii. Scientific Reports. 2017;7:40206. doi:10.1038/srep40206.

59. Westra ER, van Houte S, Oyesiku-Blakemore S, Makin B, Broniewski JM, Best A, et al. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. Current Biology. 2015;25(8):1043–1049. doi:10.1016/j.cub.2015.01.065.