RESEARCH ARTICLE

Selective Maintenance of Multiple CRISPR Arrays Across Prokaryotes

Jake L. Weissman, William F. Fagan, and Philip L.F. Johnson*

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 28% of the prokaryotic genomes in RefSeq. We use a comparative genomics approach looking across all prokaryotes to demonstrate that on average, organisms are under selection to maintain more than one CRISPR array. Given this surprising conclusion, we consider several hypotheses concerning the source of selection and include a theoretical analysis of the possibility that a trade-off between memory span and learning speed could select for both “long-term memory” and “short-term memory” CRISPR 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. Here, we explore this apparent redundancy in the context of CRISPR*-Cas immunity.

The CRISPR-Cas immune system is a powerful defense mechanism against mobile genetic elements such as viruses and plasmids, and is the only known 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 RNA fragments that guide CRISPR-associated (Cas) proteins to degrade matching foreign DNA or RNA.^{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.

CRISPR systems appear to be widespread across diverse bacterial and archaeal lineages, with previous analyses of genomic databases indicating that 40% of bacteria and 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 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.

*Clustered Regularly Interspaced Short Palindromic Repeats.

Department of Biology, University of Maryland College Park, College Park, Maryland.

*Address correspondence to: Philip L.F. Johnson, Department of Biology, University of Maryland College Park, 1210 Biology-Psychology Building, 4094 Campus Drive, College Park, MD 20742, E-mail: plfj@umd.edu

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., Horvath *et al.*¹⁸ and Cai *et al.*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, some microbes may experience selection for multiple sets of *cas* genes or CRISPR arrays.

We suspected that prokaryotes may be under selection to maintain multiple CRISPR arrays, given that it is common for organisms across lineages to have multiple systems (as detailed below), and in some clades, these appear to be conserved over evolutionary time (e.g., Boudry *et al.*²¹ and Andersen *et al.*²²). Because microbial genomes have a deletion bias,^{23,24} we would expect extraneous systems to be removed over time. Here, we constructed 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 by approximately two arrays. 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 trade-off 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

Data set

All available completely sequenced prokaryotic genomes (all assembly levels, bacteria, and archaea) were downloaded from the National Center for Biotechnology Information (NCBI) non-redundant RefSeq database FTP site²⁵ on December 23, 2017. Genomes were scanned for the presence of CRISPR arrays using the CRISPRDetect v2.2.²⁶ 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. The presence or absence of *cas* genes was 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 species. In this way, we

only examined genomes known to be compatible with CRISPR immunity.

Test for selection maintaining multiple arrays

We detected selection by comparing nonfunctional (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 used the presence of both genes on a genome as a marker for functionality of arrays on that genome and the absence of one or both genes as a marker for nonfunctionality (validated in Supplementary Text S1; Supplementary Data are available online at www.liebertpub.com/crispr). This differentiation allowed us to consider the probability distributions of the number of CRISPR arrays *i* in nonfunctional (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 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 nonfunctional distribution will look like the functional distribution shifted by one (S_i):

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

for $i \geq 0$ (S_i renormalized to account for loss of 0-array category).

We take two approaches to testing this hypothesis: one parametric from first principles and one nonparametric with less power but fewer assumptions. In our parametric approach, we construct a stochastic model of neutral array accumulation and find that both N_i and S_i should fit a negative binomial distribution at equilibrium (see Supplementary Text S2 for derivation). We calculate point maximum likelihood estimates of the means of these fitted distributions ($\hat{\mu}_N$ and $\hat{\mu}_S$). We expect that $\hat{\mu}_S > \hat{\mu}_N$ if more than one array is selectively maintained, and we bootstrap confidence intervals on these estimates by resampling with replacement from our functional and nonfunctional array count distributions in order to determine whether the effect is significant.

We also construct a nonparametric test for selection by determining at what shift *s* the mismatch between $F_{i+s} / \sum_{j=s}^{\infty} F_j$ and N_i , measured as the sum of squared differences between the distributions, is minimized:

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

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

We note that the array accumulation process underlying these methods assumes that CRISPR arrays are primarily lost all at once (e.g., due to recombination between flanking insertion sequences^{28,29}) rather than through a process of gradual decay due to spacer loss. Experimental evidence supports spontaneous loss of the entire CRISPR array,³⁰ as do comparisons between closely related genomes.²⁹ We discuss this assumption and provide evidence supporting spontaneous loss in Supplementary Text S3.

CRISPR spacer turnover model

We developed 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; Supplementary Text S1 and Supplementary Table S1). This model allowed us to specify the strength of priming (i.e., if a CRISPR array has a spacer targeting a particular viral species, the rate of spacer acquisition toward that species is increased^{31,32}) and a functional form for spacer loss over time.

Using this model, we can determine the optimal spacer acquisition rate, given a particular pattern of pathogen recurrence in the environment. If the optima for distinct recurrence patterns do not overlap, it indicates that multiple arrays would be required to combat viral species simultaneously with these distinct recurrence patterns. For model analysis, see Supplementary Text S4.

We consider two functional forms for spacer loss based on known features of CRISPR biology. (1) The rate of per-spacer loss increases linearly with locus length. This form is based on the observation that spacer loss appears to occur via homologous recombination between repeats,^{33–35} which becomes more likely with increasing numbers of spacers (and thus repeats). (2) The length of an array is capped at some fixed “effective” number of spacers. This form is based on 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 are from the first few spacers^{36–38}). We analyze both models (Supplementary Text S4), though they give qualitatively similar results, and so we focus on case 1 in the Results.

Results

Having more than one CRISPR array is common

Almost half of the prokaryotic genomes in the RefSeq database have at least one CRISPR array, and around a quarter have multiple CRISPR arrays (Table 1). In contrast to this result, having more than one set of *cas* targeting genes is not nearly as common. We counted the number of signature targeting genes diagnostic for type I, II, and III systems in each genome (*cas3*, *cas9*, and *cas10*, respectively³⁹). Only 5% of all genomes have more than one targeting gene. Of these cases, about half correspond to cases of multiple types of targeting genes in the same genome (Table 1).

Some species are overrepresented in RefSeq (e.g., because of medical relevance), and we wanted to avoid results being driven by just those few particular species. We controlled for this bias by randomly subsampling 10 genomes from each species with >10 genomes in the database and found broadly similar results (Supplementary Table S4).

Selection maintains multiple CRISPR arrays

We leveraged the difference between functional and nonfunctional genomes, within each of which the process of CRISPR array accumulation should be distinct (Fig. 1 and Supplementary Table S2). Nonfunctional CRISPR arrays should accumulate neutrally in a genome following background rates of horizontal gene transfer (HGT) and gene loss (see Methods). 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 nonfunctional 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 selection maintains multiple (functional) arrays, 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 ($\Delta\mu = \hat{\mu}_S - \hat{\mu}_N = 1.09 \pm 0.03$). We bootstrapped 95% confidence intervals of our estimates by resampling genomes (Supplementary Table S2) and found that the bootstrapped distributions did not overlap, indicating a highly

Table 1. CRISPR Array and *Cas* Multiplicity Across Prokaryotic Genomes

Genome set	With CRISPR array	> 1 CRISPR arrays	> 1 signature <i>cas</i> genes	> 1 type of signature <i>cas</i> gene
Full data set	44%	28%	5%	2%
Subsampled	40%	24%	9%	5%

FIG. 1. Selection maintains more than one CRISPR array on average across prokaryotes. **(a and b)** Distribution of number of arrays per genome in **(a)** genomes with nonfunctional CRISPR immunity and **(b)** genomes with putatively functional CRISPR immunity. The tails of these distributions are cut off for ease of visual comparison (24 genomes with >10 arrays in **(a)** and 498 genomes with >10 arrays in **(b)**). In **(a)**, the black circles show the negative binomial fit to the distribution of arrays in nonfunctional genomes. In **(b)**, the black circles indicate the negative binomial fit to the single-shifted distribution ($s=1$), and the pink triangles to the double-shifted distribution ($s=2$). Note that the fit to the double-shifted distribution (pink triangles in **(b)**) visually resembles the distribution of nonfunctional arrays shown in **(a)**. **(c)** We formally quantify the difference between the nonfunctional/shifted function distributions and find an optimal shift of $s^*=2$. **(d)** The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show no overlap with 1,000 bootstrap replicates.

significant result (Fig. 1d). To control for the possibility that multiple sets of *cas* genes in a small subset of genomes could be driving this selective signature, we restricted our data set only to genomes with one or fewer signature targeting genes (*cas3*, *cas9*, or *cas10*^{12,39}) and one or fewer copies each of the genes necessary for spacer acquisition (*cas1* and *cas2*). Even in this restricted set, selection maintains more than one (functional) CRISPR array, though the effect size is smaller ($\Delta\mu=0.61 \pm 0.02$; Supplementary Fig. S1).

In order to confirm our results further, we (1) subsampled overrepresented taxa in the data set, (2) per-

formed phylogenetically corrected tests to account for possible evolutionary correlation in rates of HGT, (3) considered the effects of potential physical linkage between *cas* genes and CRISPR arrays, (4) looked for artifacts as a factor of genome assembly level, (5) considered the potential effects of CRISPR immunity on rates of HGT,⁴⁰ and, finally, (6) merged arrays with identical repeats to account for the potential formation of neo-CRISPR arrays by off-target spacer integration,⁴¹ as well as other array duplication events. In all cases, our qualitative result of selection ($\Delta\mu>0$) holds (Supplementary

Texts S5 and S6). Additionally, we explored the possibility that the CRISPR detection algorithm we used could be biased and/or suffering from a high rate of false-positives, and found our qualitative result did not change when using a higher score cutoff, restricting to arrays with experimentally verified repeat sequences, or using an alternative algorithm (Supplementary Text S7).

A trade-off between memory span and acquisition rate could select for multiple arrays in a genome

We built a simple model of spacer turnover dynamics in a single CRISPR array. We consider three patterns of viral residency in the environment, corresponding to the major threats prokaryotes are likely to face: (1) “background” viruses that coexist with their hosts over long time periods,⁴² (2) periodic outbreaks of a particular

“transient” virus that enters and leaves the system,⁴³ and (3) “novel” viruses that a host has not previously encountered (see Methods and Supplementary Text S4). For very high spacer acquisition rates, a host will be able to defend effectively against all three types of viral species simultaneously because the acquisition of immunity will be nearly instantaneous (“short-term memory”/“fast-learning” in Fig. 2 and Supplementary Fig. S2). Such high rates are unrealistic due to physical constraints on the speed of adaptation, as well as the evolutionary constraint of autoimmunity (Supplementary Text S8^{44–48}). CRISPR adaptation is rapid, but it is not instantaneous, and infected but susceptible hosts will often perish before a spacer can be acquired.⁴⁹ This is precisely why the memory-like quality of CRISPR immunity is advantageous.

FIG. 2. Immune memory is maximized at intermediate and low spacer acquisition rates, creating a trade-off 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 toward both viral species was maintained. The green region indicates where immune memory was lost toward 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 toward the background species was maintained (i.e., immune memory was rapidly lost and $t_i > 10^{-5}$). Dark blue indicates where equilibrium spacer content toward one or both species did not exceed one, despite both species being present in the system (Supplementary Text S4). **(b)** The trade-off between memory span and learning speed. The speed of immune response to the transient virus is plotted against the speed of response to a novel virus to which the system has not been previously exposed (so that there are no spacers targeting this virus), over a range of spacer acquisition rates ($\mu_A \in [10^{-3.5}, 1]$), and letting the densities of transient and background viruses be equal. The speed of immune response to a virus is defined as $1/(1+t)$ where t is time to first spacer acquisition ($t=0$ if memory is maintained). The speed of response to the novel virus is therefore $1/(1+t_N)$ where t_N is the time to first spacer acquisition toward this virus. For specifics on calculating t_i and t_N , see Supplementary Text S4. Note that μ_A is the number of spacers expected to be acquired per viral particle adsorbed to the host cell.

Our analysis also reveals a region of parameter space with low spacer acquisition rates in which immunity is maintained toward both background and transient viruses (“long-term memory”/“slow-learning” in 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 reacquire immunity quickly toward the transient virus. This sets up a trade-off between the ability of a host to defend against both transient and novel viruses, since the response time toward novel threats in the “long-term memory”/“slow-learning” region of parameter space is slow (Fig. 2b), but memory of transient threats is lost if spacer acquisition rates are increased. Thus, in order to maximize novel spacer acquisition and memory span simultaneously, a two-system solution will be required.

Additionally, priming expands the “washout” region of parameter space because high spacer uptake from background viruses will crowd out long-term immune memory (Supplementary Fig. S3). This suggests that priming strengthens the learning versus memory trade-off and makes a two-array solution more likely.

Selection varies between taxa and system types

A handful of species in the data set were represented by a large number of genomes (>1,000), with at least one each of functional and nonfunctional genomes. We performed our test for selection on each of these species individually and found a large amount of variation between species (Supplementary Table S3). Notably, genomes of *Campylobacter jejuni*, *Escherichia coli*, and *Salmonella enterica* show evidence for selection against having a functional CRISPR array (negative $\Delta\mu$), indicating that CRISPR immunity is selected against on average in some groups of organisms. Previous work has shown that CRISPR in *E. coli* and *S. enterica* appears to be nonfunctional as an immune system under natural conditions.^{50,51} We had relatively few archaeal genomes (<1% of the data set), but they showed a clear signal of selection maintaining multiple arrays ($\Delta\mu=1.05\pm0.56$; Supplementary Fig. S4).

While we do not have direct information on system type for the majority of arrays in our data set, we can subdivide genomes into those containing the signature *cas* targeting genes for types I, II, or III CRISPR systems (*cas3*, *cas9*, and *cas10*, respectively) as a proxy for system type.³⁹ The number of arrays per genome differed significantly among system types (Supplementary Fig. S5), and the largest difference was between genomes with class I targeting proteins, which had around two arrays on average (type I and type III, 2.10 and 1.96, respectively), and

class II targeting proteins, which only had one array on average (type II, 1.05). We excluded genomes with multiple types of targeting genes from this analysis.

We could not run our test for selection directly on these subsets of the data, since they exclude genomes without arrays or *cas* genes. Instead, we classified species into types if the only observed targeting gene type among all representatives of that species corresponded to a particular type. Thus, we could test for our signature of selection among species that “favor” a particular type of CRISPR system. All types showed a signature of multi-array selection ($\Delta\mu=1.09\pm0.05$, 0.62 ± 0.02 , and 1.79 ± 0.06 , respectively). In particular, type III “species” had an exceptionally strong signal, and organisms in this group may be under selection to maintain three arrays.

Discussion

Selection maintains multiple CRISPR arrays across prokaryotes

On average, prokaryotes are under selection to maintain more than one CRISPR array. The number of CRISPR arrays in a genome appears to follow a negative binomial distribution quite well (Fig 1 and Supplementary Figs. S1, S6, and S7), consistent with our theoretical prediction. We note that due to the large size of this data set, 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 selection is conservative to the miscategorization of arrays as “functional” or “nonfunctional.” Miscategorizations could occur for several reasons because preexisting spacers may continue to confer immunity, some CRISPR arrays may be conserved for nonimmune purposes (e.g., Touchon and Rocha⁵⁰ and Li *et al.*⁵²), and intact acquisition machinery is no guarantee of system functionality. Our test is conservative precisely because of such miscategorizations, as they should drive $\hat{\mu}_N$ and $\hat{\mu}_S$ closer to each other. Selection against having a CRISPR array in nonfunctional genomes could produce a false signature of multi-array selection, but this is unlikely because nonfunctional arrays probably carry extremely low or nonexistent associated costs.⁵³

Our test for selection is also robust to false-positive or -negative array discovery rates because it relies on relative differences between array counts in functional and nonfunctional genomes, not their absolute values. The only problem could arise if the discovery error rates were different between the two categories. However, the array detection process did not take functionality into account, and we found only a marginal difference in CRISPRDetect confidence scores between the two groups (Supplementary Fig. S8). We further confirmed this robustness to

peculiarities of the detection algorithm by changing our CRISPRDetect score threshold and comparing to the distribution of arrays per genome in the independently generated CRISPR database (Supplementary Text S7⁵⁴).

Finally, we note that $\hat{\mu}_N$ and $\hat{\mu}_S$ take on a range of values, depending on what subset of taxa/genomes is considered. This is to be expected, as each set of species will occupy a distinct environment in terms of both the rate of HGT and the usefulness of CRISPR immunity. Nevertheless, our qualitative signature of selection is robust to this quantitative variability.

Why have multiple CRISPR-Cas systems?

Possibly, multiple arrays could be selectively maintained, even in the absence of any fitness advantage, if, by chance, each array acquired complementary spacer content toward distinct viral targets. In type I and type II systems, if arrays share acquisition machinery, then such complementarity is unlikely because priming will ensure both arrays contain spacers toward any target encountered, meaning that the content of the two arrays will be largely redundant.³¹ Type III systems are unprimed and have slow spacer acquisition rates,⁵⁵ and therefore may be maintained via spacer complementarity, perhaps explaining why species favoring type III systems appear to experience selection maintaining three rather than just two CRISPR arrays. Even in type I and type II systems, if each array is associated with a separate set of spacer acquisition machinery, then cross-priming will be less likely, and complementarity could arise. Nevertheless, this does not explain the multi-array conservation we see in genomes with only a single set of *cas* genes.

Therefore, we are left with two broadly defined reasons why having multiple CRISPR arrays might be adaptive: (1) multiple similar systems could lead to improved immunity through redundancy, and (2) multiple dissimilar systems could allow specialization toward distinct types of threats.

In the case of similar systems, immunity could be improved by (1) an increased spacer acquisition rate, (2) an increased rate of targeting, or (3) a longer time to expected loss of immunity. Duplication of *cas* genes could increase uptake (1) and targeting rates (2), but again this could not explain our results with a single set of *cas* genes. Alternatively, duplication of CRISPR arrays could increase targeting (2) by producing a larger number of crRNA transcripts or increase memory duration (3) through spacer redundancy. However, the effectiveness of crRNA may actually decrease in the presence of competing crRNAs,^{53,56,57} and spacer redundancy across multiple arrays has little advantage over redundancy within a single array (Supplementary Text S9).

On a larger scale, redundancy of either arrays or *cas* genes might be a form of bet-hedging against mutation-induced loss of functionality of the CRISPR system.^{30,58}

Alternatively, dissimilar systems could help defend against diverse threats. Diverse *cas* genes may allow hosts to evade broadly acting anti-CRISPR proteins encoded by some viruses.^{59,60} Indeed, promiscuous type III Cas proteins are often encoded alongside type I systems and can cooperate to target phages that have mutated to escape type I targeting.⁶¹ Empirically, we see the inclusion of genomes with multiple *cas* targeting genes increases the effect size of our test for selection, suggesting these factors may play a role. However, these *cas*-diversity hypotheses cannot explain the signature for multi-array adaptiveness observed among genomes with only a single set of targeting proteins. Note that we observed our signature of selection on multiple arrays when limiting our analyses to arrays with both identical (Supplementary Text S10) and dissimilar (Supplementary Text S6) repeat sequences. Therefore, selective maintenance of multiple arrays does not appear to be isolated to genomes with arrays of the same type or different types, but rather seems to be a much more general phenomenon. Additionally, given the very small number of genomes with multiple types of *cas* targeting genes in our data set, it is unlikely that selection for multiple types of systems is particularly widespread, even if it does exist in some cases.

We developed a hypothesis that diversity in spacer acquisition rate among arrays could lead to selection for multiple arrays. Our theoretical model illustrates how factors intrinsic to the mechanism of CRISPR immunity could create a trade-off between memory span and learning speed. Either the physical loss of spacers due to homologous recombination or the effective loss of spacers due to differential transcription along the array leads to a qualitatively similar result. In both cases, rapid spacer uptake causes rapid spacer loss (either physical or effective), producing the aforementioned trade-off. A low acquisition rate system is unlikely to pick up a spacer from a single viral exposure, but over a long time frame, it may acquire spacers from viruses that periodically reappear in the system. Additionally, recombination between arrays⁶² could potentially facilitate the passage of memories between “fast” and “slow” arrays, allowing short-term memories to become long-term ones.

While we do not have empirical evidence that rate variation drives the observed signature of selection of multiple arrays, this hypothesis remains attractive, since it can explain the signature, even in the absence of multiple sets of *cas* genes. Acquisition rates vary between arrays, even on the same genome^{42,63} and even when those arrays

share *cas* genes and have an identical or nearly identical repeat sequence.^{64,65} We found no clear link between the diversity of repeat sequences and a proxy for spacer acquisition rates (Supplementary Text S10). Further, we found indications of selection, even when restricting to arrays with identical repeats (Supplementary Text S10). Thus, the factors influencing acquisition rate appear to be idiosyncratic, perhaps related to the genomic position of the CRISPR array.

When partial spacer-target matches exist, variability in spacer acquisition rates among arrays will be largely irrelevant because priming will ensure rapid acquisition of new spacers. On the other hand, when no match exists, due to either spacer loss or the introduction of a truly novel viral species into the environment, primed spacer uptake will not occur. Thus, the rate at which a host encounters novel threats will determine the importance of the baseline spacer acquisition rate. In environments where novel viruses are frequently encountered, small differences in acquisition rate can be important for host fitness, whereas in environments where host and virus pairs consistently coevolve over time, priming will be the more important phenomenon.

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., Iranzo *et al.*,^{4,16} Kumar *et al.*,⁴⁵ and Westra *et al.*⁶³), 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

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

Author Disclosure Statement

No competing financial interests exist.

References

- Makarova KS, Wolf YI, Snir S, et al. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol* 2011;193:6039–6056. DOI: 10.1128/JB.05535-11.
- Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. *Nucl Acids Res* 2013;41:4360–4377. DOI: 10.1093/nar/gkt157.
- van Houte S, Buckling A, Westra ER. Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol Mol Biol Rev* 2016;80:745–763. DOI: 10.1128/MMBR.00011-16.
- Iranzo J, Lobkovsky AE, Wolf YI, et al. 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.
- Makarova KS, Grishin NV, Shabalina SA, et al. 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. *Biol Direct* 2006;1:7. DOI: 10.1186/1745-6150-1-7.
- Goren M, Yosef I, Edgar R, et al. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA Biol* 2012;9:549–554. DOI: 10.4161/rna.20177.
- Mojica FJM, Díez-Villaseñor C, García-Martínez J, et al. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 2005;60:174–182. DOI: 10.1007/s00239-004-0046-3.
- Bolotin A, Quinquis B, Sorokin A, et al. Clustered regularly interspaced short palindromic repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 2005;151:2551–2561. DOI: 10.1099/mic.0.28048-0.
- Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;315:1709–1712. DOI: 10.1126/science.1138140.
- Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 2008;322:1843–1845. DOI: 10.1126/science.1165771.
- Marraffini LA. CRISPR-Cas immunity in prokaryotes. *Nature* 2015;526:55–61. DOI: 10.1038/nature15386.
- Makarova KS, Haft DH, Barrangou R, et al. Evolution and classification of the CRISPR–Cas systems. *Nat Rev Microbiol* 2011;9:467–477. DOI: 10.1038/nrmicro2577.
- Sorek R, Lawrence CM, Wiedenheft B. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem* 2013;82:237–266. DOI: 10.1146/annurev-biochem-072911-172315.
- Burstein D, Harrington LB, Strutt SC, et al. New CRISPR–Cas systems from uncultivated microbes. *Nature* 2017;542:237–241. DOI: 10.1038/nature21059.
- Burstein D, Sun CL, Brown CT. Major bacterial lineages are essentially devoid of CRISPR–Cas viral defence systems. *Nat Commun* 2016;7:10613. DOI: 10.1038/ncomms10613.
- Iranzo J, Lobkovsky AE, Wolf YI, et al. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR–Cas in an explicit ecological context. *J Bacteriol* 2013;195:3834–3844. DOI: 10.1128/JB.00412-13.
- Weinberger AD, Wolf YI, Lobkovsky AE, et al. Viral diversity threshold for adaptive immunity in prokaryotes. *MBio* 2012;3:e00456–12. DOI: 10.1128/mBio.00456-12.
- Horvath P, Couët-Monvoisin A-C, Romero DA, et al. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* 2009;131:62–70. DOI: 10.1016/j.ijfoodmicro.2008.05.030.
- Cai F, Axen SD, Kerfeld CA. Evidence for the widespread distribution of CRISPR–Cas system in the Phylum Cyanobacteria. *RNA Biol* 2013;10:687–693. DOI: 10.4161/rna.24571.
- Puigbò P, Makarova KS, Kristensen DM, et al. Reconstruction of the evolution of microbial defense systems. *BMC Evol Biol* 2017;17:94. DOI: 10.1186/s12862-017-0942-y.
- Boudry P, Semenova E, Monot M, et al. Function of the CRISPR–Cas system of the human pathogen *Clostridium difficile*. *MBio* 2015;6:e01112–15. DOI: 10.1128/mBio.01112-15.
- Andersen JM, Shoup M, Robinson C, et al. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biol Evol* 2016;8:2841–2855. DOI: 10.1093/gbe/evw203.
- Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet* 2001;17:589–596.
- Kuo C-H, Ochman H. Deletional bias across the three domains of life. *Genome Biol Evol* 2009;1:145–152. DOI: 10.1093/gbe/evp016.
- O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl Acids Res* 2016;44:D733–D745. DOI: 10.1093/nar/gkv1189.
- Biswas A, Staats RH, Morales SE, et al. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genom* 2016;17:356. DOI: 10.1186/s12864-016-2627-0.
- Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. *Nucl Acids Res* 2016;44:6614–6624. DOI: 10.1093/nar/gkw569.

28. Almendros C, Mojica FJ, Díez-Villaseñor C, et al. CRISPR-Cas functional module exchange in *Escherichia coli*. *MBio* 2014;5:e00767–13.
29. Shah SA, Garrett RA. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res Microbiol* 2011;162:27–38. DOI: 10.1016/j.resmic.2010.09.001.
30. Jiang W, Maniv I, Arain F, et al. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet* 2013;9:e1003844. DOI: 10.1371/journal.pgen.1003844.
31. Datsenko KA, Pougach K, Tikhonov A, et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* 2012;3:945. DOI: 10.1038/ncomms1937.
32. Swarts DC, Mosterd C, van Passel MWJ, et al. CRISPR interference directs strand specific spacer acquisition. *PLoS One* 2012;7:e35888. DOI: 10.1371/journal.pone.0035888.
33. Garrett RA, Shah SA, Vestergaard G, et al. CRISPR-based immune systems of the *Sulfolobales*: complexity and diversity. *Biochem Soc Trans* 2011;39:51–57. DOI: 10.1042/BST0390051.
34. Gudbergsson S, Deng L, Chen Z, et al. Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol Microbiol* 2011;79:35–49. DOI: 10.1111/j.1365-2958.2010.07452.x.
35. Weinberger AD, Sun CL, Pluciński MM, et al. Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput Biol* 2012;8:e1002475. DOI: 10.1371/journal.pcbi.1002475.
36. Bernick DL, Cox CL, Dennis PP, et al. Comparative genomic and transcriptional analyses of CRISPR systems across the genus *Pyrobaculum*. *Front Microbiol* 2012;3. DOI: 10.3389/fmicb.2012.00251.
37. Hale CR, Majumdar S, Elmore J, et al. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell* 2012;45:292–302. DOI: 10.1016/j.molcel.2011.10.023.
38. Richter H, Zoephel J, Schermuly J, et al. Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucl Acids Res* 2012;40:9887–9896. DOI: 10.1093/nar/gks737.
39. Makarova KS, Wolf YI, Alkhnbashi OS, et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 2015;13:722–736. DOI: 10.1038/nrmicro3569.
40. Watson BN, Staals RH, Fineran PC. CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio* 2018;9:e02406–17. DOI: 10.1128/mBio.02406-17.
41. Nivala J, Shipman SL, Church GM. Spontaneous CRISPR loci generation *in vivo* by non-canonical spacer integration. *Nat Microbiol* 2018;3:310–318. DOI: 10.1038/s41564-017-0097-z.
42. Paez-Espino D, Sharon I, Morovic W, et al. CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *MBio* 2015;6:e00262–15. DOI: 10.1128/mBio.00262-15.
43. Chow C-ET, Fuhrman JA. Seasonality and monthly dynamics of marine myovirus communities. *Environ Microbiol* 2012;14:2171–2183. DOI: 10.1111/j.1462-2920.2012.02744.x.
44. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes Dev* 2015;29:356–361. DOI: 10.1101/gad.257550.114.
45. Kumar MS, Plotkin JB, Hannehalli S. Regulated CRISPR modules exploit a dual defense strategy of restriction and abortive infection in a model of prokaryote-phage coevolution. *PLoS Comput Biol* 2015;11:e1004603. DOI: 10.1371/journal.pcbi.1004603.
46. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucl Acids Res* 2012;40:5569–5576. DOI: 10.1093/nar/gks216.
47. Levy A, Goren MG, Yosef I, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 2015;520:505–510. DOI: 10.1038/nature14302.
48. Stern A, Keren L, Wurtzel O, et al. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* 2010;26:335–340. DOI: 10.1016/j.tig.2010.05.008.
49. Hynes AP, Villion M, Moineau S. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nat Commun* 2014;5:4399. DOI: 10.1038/ncomms5399.
50. Touchon M, Rocha EPC. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One* 2010;5:e11126. DOI: 10.1371/journal.pone.0011126.
51. Touchon M, Charpentier S, Clermont O, et al. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* 2011;193:2460–2467. DOI: 10.1128/JB.01307-10.
52. Li R, Fang L, Tan S, et al. Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Res* 2016;26:1273–1287. DOI: 10.1038/cr.2016.135.
53. Martynov A, Severinov K, Isolatov I. Optimal number of spacers in CRISPR arrays. *PLoS Comput Biol* 2017;13:e1005891.
54. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 2007;8:172. DOI: 10.1186/1471-2105-8-172.
55. Pyenson NC, Gayvert K, Varble A, et al. Broad targeting specificity during bacterial type III CRISPR-Cas immunity constrains viral escape. *Cell Host Microbe* 2017;22:343–353. DOI: 10.1016/j.chom.2017.07.016.
56. Stachler A-E, Marchfelder A. Gene Repression in Haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas I-B system. *J Biol Chem* 2016;291:15226–15242. DOI: 10.1074/jbc.M116.724062.
57. Stachler A-E, Turgeman-Grott I, Shtifman-Segal E, et al. High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. *Nucl Acids Res* 2017;45:5208–5216. DOI: 10.1093/nar/gkx150.
58. Weissman JL, Holmes R, Barrangou R, et al. Immune loss as a driver of coexistence during host-phage coevolution. *ISME J* 2018;12:585–597. DOI: 10.1038/ismej.2017.194.
59. Bondy-Denomy J, Pawluk A, Maxwell KL, et al. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* 2013;493:429–432. DOI: 10.1038/nature11723.
60. Pawluk A, Bondy-Denomy J, Cheung VHW, et al. A new group of phage anti-CRISPR gene inhibits the type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *MBio* 2014;5:e00896-14. DOI: 10.1128/mBio.00896-14.
61. Silas S, Lucas-Elio P, Jackson SA, 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.
62. Kupczok A, Landan G, Dagan T. The contribution of genetic recombination to CRISPR array evolution. *Genome Biol Evol* 2015;7:1925–1939. DOI: 10.1093/gbe/evv113.
63. Westra ER, van Houte S, Oyesiku-Blakemore S, et al. Parasite exposure drives selective evolution of constitutive versus inducible defense. *Curr Biol* 2015;25:1043–1049. DOI: 10.1016/j.cub.2015.01.065.
64. Staals RHJ, Jackson SA, Biswas A, et al. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun* 2016;7:12853. DOI: 10.1038/ncomms12853.
65. Zeng H, Zhang J, Li C, et al. The driving force of prophages and CRISPR-Cas system in the evolution of *Cronobacter sakazakii*. *Sci Rep* 2017;7:40206. DOI: 10.1038/srep40206.

Received for publication August 17, 2018;
 Revised October 15, 2018;
 Accepted November 6, 2018.