Is having more than one CRISPR array adaptive?

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

July 26, 2017

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. This adaptive signature appears to be a function of more subtle diversity between the CRISPR systems rather than their multiplicity alone. We go on to develop a mathematical model of CRISPR immune memory turnover to show how a tradeoff between memory span and learning speed can lead to selection for "long-term memory" and "short-term memory" arrays.

Significance Statement

Many viruses infect bacteria and archaea. To protect themselves, these microbes employ a variety of defense mechanisms. Surprisingly, many microbes have multiple copies of the same type of defense system encoded on their genome. We determine whether this apparent redundancy of immunity is adaptive, and why it would be so. Specifically, we examine the CRISPR immune system, which allows microbes to store immune "memories" of past infections and use these memories to fight future infections. Using publicly available genomic data we show that having two CRISPR systems is adaptive, on average. We also build a theoretical model indicating it is adaptive for microbes to have CRISPR systems specializing as both "short-term" and "long-term" memory for rapid response and reliable storage.

1 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 [29, 28, 16]. While having multiple types of immune systems may decrease the chance of pathogen evolutionary escape [18], 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 [25, 13]. 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 [33, 4, 2]. 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 [2, 31, 30]. 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 [26, 38, 7]. 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 [2, 26]. 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 [8]. Other studies suggest that particular environmental factors can be important in determining whether or not CRISPR immunity is effective (e.g., in thermophilic environments [17, 49]). 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., [15, 9]). CRISPR and other immune systems are horizontally transferred at a high rate relative to other genes in bacteria [36], 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. [6, 1]). Since microbial genomes have a deletion bias [32, 21], we would expect extraneous systems to be removed over time. Here we use publicly available genome data to provide the first large-scale evidence that selection actively maintains more than one CRISPR array in a wide range of bacteria and archea. We then go on to test several hypotheses for why having multiple arrays might be adaptive., using both comparative genomics and theoretical approaches. We conclude 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.

2 Methods

2.1 Dataset

All available prokaryotic sequences were downloaded from NCBI's non-redundant RefSeq database FTP site (`ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria`, [34]) on May 11, 2017. Genomes were scanned for the presence of CRISPR arrays using the CRISPRDetect software [3]. 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 [44]. We discarded genomes without *cas1* and *cas2* 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.

2.2 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. 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. This hypothesis predicts that the non-functional distribution

will look like the functional distribution shifted by one: $N_i = F_{i+1} / \sum_{j=1}^{\infty} F_j$ 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) [36]. 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 both the non-functional data and for the functional data as shifted under our null hypothesis ($S_i = F_{i+1} / \sum_{i=1}^{\infty} F_i$). 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 is unclear how this will be reflected in these parameters. Therefore it is more useful to compare the means of the distributions $\mu_k = \frac{p_k r_k}{1-p_k}$, $k \in N, S$. 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:

$$s^\star = \underset{s}{\text{argmin}} \sum_{i=0}^{\infty} \left(N_i - F_{i+s} / \sum_{j=s}^{\infty} F_j \right)^2 . \tag{1}$$

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.

2.3 Correcting for correlations in HGT

Differential rates of horizontal gene transfer (HGT) between lineages could produce an observed correlation between *cas* presence and array count in the absence of any selection for having multiple CRISPR arrays. In other words, some lineages would have *cas* genes and many arrays due to a high arrival rate of foreign genetic material, and other lineages would lack *cas* genes and 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 *cas*-possessing and *cas*-lacking 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.

2.4 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):

$$\underbrace{\frac{dC_i}{dt}}_{\text{Spacers Targeting Viral Spp. } i} = \underbrace{a_i(t, C_i)}_{\text{Acquisition}} - \underbrace{\mu_L C_i \sum_j C_j}_{\text{Loss}} \tag{2}$$

where μ_L is the spacer loss rate parameter and a_i will be 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 [11, 14, 48], 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 "fluctuating" species (F) that leaves and returns to the environment after some interval of time ($f_F(t)$ is a binary function that takes a value of one if virus F is present in the environment and zero otherwise).

We also can consider the phenomenon of priming in our model, wherein if a CRISPR system has a spacer targeting a particular viral species, the rate of spacer acquisition towards that species is increased [10, 43]. Thus

$$a_i(t, C_i) = \mu_A v_i f_i(t) g(C_i) \tag{3}$$

where

$$g(C_i) = \begin{cases} 1 & C_i < 1 \\ p & C_i \geq 1 \end{cases} \tag{4}$$

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.

3 Results

3.1 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 [27]). 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%).

3.2 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 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, [12]). 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 [22] and arrays may have been inherited recently from strains with active *cas* machinery.

3.3 Having more than one CRISPR array is adaptive

We leveraged the difference between genomes that possessed or lacked *cas* spacer acquisition machinery (*cas1* and *cas2*, Fig. 1, Table 1). Without *cas1* and *cas2*, CRISPR arrays will be non-functional and 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 *cas*-lacking genomes ($\hat{\mu}_N$, Fig. 1a). The other came from the *cas*-possessing 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* [26, 27]) 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

			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.

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 (Section 2.2). First, beyond the clear effect of the presence of *cas* genes on the number of arrays in a genome, we do not see that an increased number of *cas1* genes in a genome has 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 *cas*-possessing and *cas*-lacking 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 (Section 2.2) that controls for differences in rates of HGT between lineages also confirms a signature of multi-array adaptiveness, though the effect is less strong ($\bar{\Delta\mu}_k = 0.44 \pm 0.14$). 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.

3.4 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 (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 drawn repeat sequences except for a few conserved sites between them (S7D Fig). We also observed that among genomes with cas genes present, the area of the peak corresponding to dissimilar repeat sequences was significantly higher than among genomes lacking *cas* genes ($\chi^2 = 16.784$, $df = 1$, $p < 4.19 \times 10^{-5}$, S7 Fig). This suggests that the observed signature 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

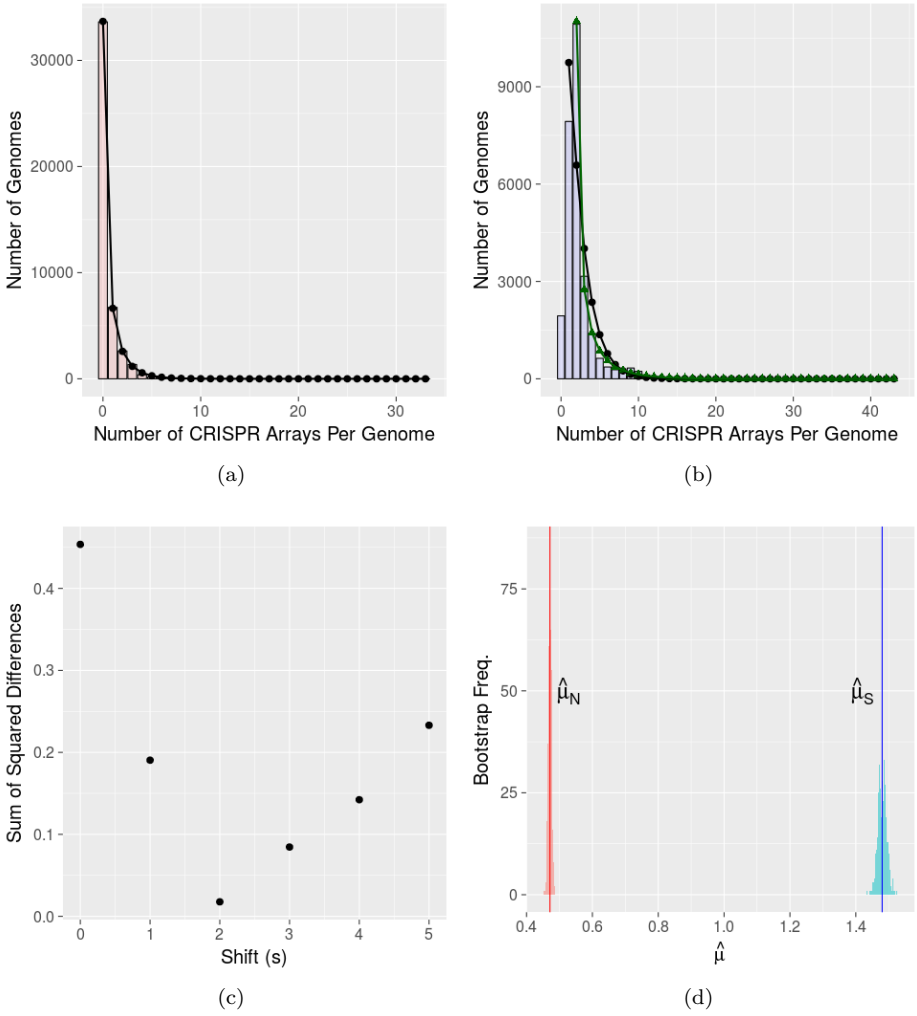

Figure 1: Having more than one CRISPR array is adaptive on average across prokaryotes. (a-b) Distribution of number of arrays per genome in (a) genomes that lacked *cas1*, *cas2*, or both, and (b) genomes that had *cas1* and *cas2* genes. In (a) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and green triangles to the double-shifted distribution ($s = 2$). In (b) the black circles show the negative binomial fit to the distribution of arrays in *cas*-lacking genomes. (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.

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 degree of variability in array consensus repeat sequences within a genome was associated with variability in array length, measured as 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. This relationship had poor predictive power, but was significant ($R^2 = 0.007464$, $p < 0.00123$). The relationship was not driven by genomes with extremely low or high length-variable arrays (top and bottom 5% excluded, $R^2 = 0.01041$, $p < 0.000698$).

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

The evidence in Section 3.4 suggests that multi-array adaptiveness is linked to differences in consensus repeat sequences between arrays and that these differences may be associated with the spacer acquisition rate of each array. 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. 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. 2, S9 Fig). We found that for very high spacer acquisition rates, the system is able to maintain immunity to both background and fluctuating viral populations. High rates of spacer acquisition are unrealistic as they lead to high rates of autoimmunity ([47, 20, 51, 23, 42], S2 Text). Our analysis also reveals that there is a region of parameter space with low spacer acquisition rates in which immunity is maintained. This is the region where low spacer turnover rates allow immune memory to remain in the system over longer periods of time (Fig. 2b). In contrast to this result, if we examine the time to first spacer acquisition when a third, novel phage species is introduced, we find that high spacer acquisition rates are favored for a quicker response to novel threats (Fig. 2b).

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 but acquisition is not rapid enough to quickly adapt to novel threats (Fig. 2b). The relative densities of the different viral species modulate the relative importance of fast-acquisition versus memory span (Fig. 2a). Thus for a range of pathogenic environments the fitness landscape is bimodal with respect to the spacer acquisition rate (taking immune maintenance as our measure of fitness). We also note that high levels of priming expand this "washout" region, as high spacer uptake from background viruses will crowd out long term immune memory (S9

Fig).

3.6 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 ($\Delta\mu = 0.60 \pm 0.06, 0.63 \pm 0.20$ respectively), though relatively few of the *S. aureus* had *cas1* and *cas2* (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 ($\Delta\mu = 0.09 \pm 0.06, 0.12 \pm 0.05$ 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 [46, 45]. 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 both CRISPR arrays and cas genes, making our approach less powerful (S10 Fig). Further, if those few genomes with non-functional CRISPR lost their *cas* machinery recently, then our power would be reduced even more because the arrays in their genomes might still bear the remnants of past selection.

4 Discussion

4.1 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 *cas*-possessing genomes are more diverse than in those without the *cas* acquisition machinery, indicating that array diversity may be important in addition to array multiplicity.

The data appear to follow a negative binomial distribution quite well (Figs

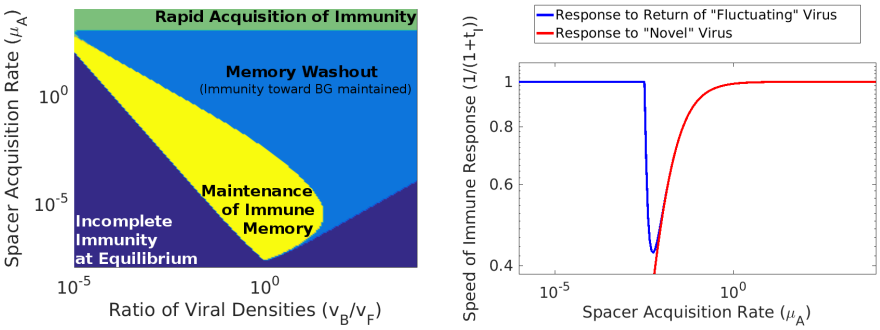

Figure 2: The optimal spacer acquisition rate with respect to continuous immunity has peaks at low and high values. (a) Phase diagram of the behavior of our CRISPR array model with two viral species, a constant "background" population and a "fluctuating" population that leaves and returns to the system at some fixed interval (Section 2.4, S1 Text). The yellow region indicates that immunity towards both viral species was maintained. The green region indicates where immune memory was lost towards the fluctuating phage species, but reacquired almost immediately upon phage reintroduction. The light blue region indicates that only immunity towards the background species was maintained (i.e., immune memory was rapidly lost). 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 results of the same model, with immunity towards the fluctuating species (blue) as in (a) and the background species present but not shown. Additionally, we have plotted the time to first spacer acquisition after the introduction of a novel phage species (red), in order to demonstrate the tradeoff between the maintenance of immune memory and the ability to respond to novel threats. Response time (t_I) is measured as the amount of time after viral infection when the first spacer targeting that virus appears in the array (zero if memory maintained).

1b and 1a, S2 Fig-S4 Fig), consistent with our theoretical prediction. This pattern is robust to subsetting 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. [46, 24]), 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.

4.2 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 [40, 41] and, since a single array can have multiple spacers with the same target, there is little advantage to having multiple arrays (S3 Text). Redundant arrays might also be a form of bet-hedging since CRISPR functionality is lost at a high rate in some prokaryotes [19]. While this last explanation is plausible, our data reveal a link between repeat diversity and functionality, which suggests that dissimilar systems play a key adaptive role.

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 [5, 35, 37]. Diverse *cas* genes may allow hosts to evade the action of these anti-CRISPR proteins, which are often extremely broadly acting

[5, 35]. 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 coevolution with anti-CRISPR proteins remains an interesting candidate to explain CRISPR multiplicity in some prokaryotes, the majority of the genomes in the dataset have only one set of cas genes and thus this this mechanism cannot explain the signature for multi-array adaptiveness observed in the majority of the dataset.

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 [39], and our data suggests a link between consensus repeat sequence and acquisition rate. Arrays with slightly different consensus repeat sequences may differ in length, despite sharing a set of *cas* genes [52]. This suggests a functional role for repeat sequence modifications in determining spacer insertion rates. 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.

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 (S11 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. [17, 18, 20, 50]), we emphasize the importance of the multiplicity of immune systems in the evolution of host defense. Here we show how a surprising amount of strategic diversity masquerades as simple redundancy.

5 Acknowledgements

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] Joakim M. Andersen, Madelyn Shoup, Cathy Robinson, Robert Britton, Katharina E. P. Olsen, and Rodolphe Barrangou. CRISPR Diversity and Microevolution in Clostridium difficile. *Genome Biology and Evolution*, 8(9):2841–2855, September 2016.

[2] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, and Philippe Horvath. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, 315(5819):1709–1712, March 2007.

[3] Ambarish Biswas, Raymond H.J. Staals, Sergio E. Morales, Peter C. Fineran, and Chris M. Brown. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, 17:356, 2016.

[4] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561, 2005.

[5] Joe Bondy-Denomy, April Pawluk, Karen L. Maxwell, and Alan R. Davidson. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, 493(7432):429–432, January 2013.

[6] Pierre Boudry, Ekaterina Semenova, Marc Monot, Kirill A. Datsenko, Anna Lopatina, Ognjen Sekulovic, Maicol Ospina-Bedoya, Louis-Charles Fortier, Konstantin Severinov, Bruno Dupuy, and Olga Soutourina. Function of the CRISPR-Cas System of the Human Pathogen Clostridium difficile. *mBio*, 6(5):e01112–15, October 2015.

[7] David Burstein, Lucas B. Harrington, Steven C. Strutt, Alexander J. Probst, Karthik Anantharaman, Brian C. Thomas, Jennifer A. Doudna, and Jillian F. Banfield. New CRISPR–Cas systems from uncultivated microbes. *Nature*, 542(7640):237–241, February 2017.

[8] David Burstein, Christine L. Sun, Christopher T. Brown, Itai Sharon, Karthik Anantharaman, Alexander J. Probst, Brian C. Thomas, and Jillian F. Banfield. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications*, 7:10613, February 2016.

[9] Fei Cai, Seth D. Axen, and Cheryl A. Kerfeld. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. *RNA Biology*, 10(5):687–693, May 2013.

[10] Kirill A. Datsenko, Ksenia Pougach, Anton Tikhonov, Barry L. Wanner, Konstantin Severinov, and Ekaterina Semenova. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*, 3:945, July 2012.

[11] Roger A. Garrett, Shiraz A. Shah, Gisle Vestergaard, Ling Deng, Soley Gudbergsdottir, Chandra S. Kenchappa, Susanne Erdmann, and Qunxin She. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochemical Society Transactions*, 39(1):51–57, February 2011.

[12] Uri Gophna, David M Kristensen, Yuri I Wolf, Ovidiu Popa, Christine Drevet, and Eugene V Koonin. No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. *The ISME Journal*, 9(9):2021–2027, September 2015.

[13] Moran Goren, Ido Yosef, Rotem Edgar, and Udi Qimron. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA biology*, 9(5):549–554, May 2012.

[14] Soley Gudbergsdottir, Ling Deng, Zhengjun Chen, Jaide V. K. Jensen, Linda R. Jensen, Qunxin She, and Roger A. Garrett. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Molecular Microbiology*, 79(1):35–49, January 2011.

[15] Philippe Horvath, Anne-Claire Coûté-Monvoisin, Dennis A. Romero, Patrick Boyaval, Christophe Fremaux, and Rodolphe Barrangou. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *International Journal of Food Microbiology*, 131(1):62–70, April 2009.

[16] Stineke van Houte, Angus Buckling, and Edze R. Westra. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiology and Molecular Biology Reviews*, 80(3):745–763, September 2016.

[17] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Evolutionary Dynamics of the Prokaryotic Adaptive Immunity System CRISPR-Cas in an Explicit Ecological Context. *Journal of Bacteriology*, 195(17):3834–3844, September 2013.

[18] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evolutionary Biology*, 15:43, 2015.

[19] Wenyan Jiang, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R. Levin, and Luciano A. Marraffini. Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. *PLoS Genet*, 9(9):e1003844, September 2013.

[20] M. Senthil Kumar, Joshua B. Plotkin, and Sridhar Hannenhalli. Regulated CRISPR Modules Exploit a Dual Defense Strategy of Restriction and Abortive Infection in a Model of Prokaryote-Phage Coevolution. *PLoS computational biology*, 11(11):e1004603, November 2015.

[21] Chih-Horng Kuo and Howard Ochman. Deletional Bias across the Three Domains of Life. *Genome Biology and Evolution*, 1:145–152, January 2009.

[22] Anne Kupczok, Giddy Landan, and Tal Dagan. The Contribution of Genetic Recombination to CRISPR Array Evolution. *Genome Biology and Evolution*, 7(7):1925–1939, July 2015.

[23] Asaf Levy, Moran G. Goren, Ido Yosef, Oren Auster, Miriam Manor, Gil Amitai, Rotem Edgar, Udi Qimron, and Rotem Sorek. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*, 520(7548):505–510, April 2015.

[24] Rongpeng Li, Lizhu Fang, Shirui Tan, Min Yu, Xuefeng Li, Sisi He, Yuquan Wei, Guoping Li, Jianxin Jiang, and Min Wu. Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Research*, 26(12):1273–1287, December 2016.

[25] Kira S. Makarova, Nick V. Grishin, Svetlana A. Shabalina, Yuri I. Wolf, and Eugene V. Koonin. 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*, 1:7, 2006.

[26] Kira S. Makarova, Daniel H. Haft, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Yuri I. Wolf, Alexander F. Yakunin, John van der Oost, and Eugene V. Koonin. Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, 9(6):467–477, June 2011.

[27] Kira S. Makarova, Yuri I. Wolf, Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen, and Eugene V. Koonin. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*, 13(11):722–736, November 2015.

[28] Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*, 41(8):4360–4377, April 2013.

[29] Kira S. Makarova, Yuri I. Wolf, Sagi Snir, and Eugene V. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of Bacteriology*, 193(21):6039–6056, November 2011.

[30] Luciano A. Marraffini. CRISPR-Cas immunity in prokaryotes. *Nature*, 526(7571):55–61, October 2015.

[31] Luciano A. Marraffini and Erik J. Sontheimer. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)*, 322(5909):1843–1845, December 2008.

[32] A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in genetics: TIG*, 17(10):589–596, October 2001.

[33] Francisco J. M. Mojica, Chc)sar Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 60(2):174–182, 2005.

[34] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(Database issue):D733–D745, January 2016.

[35] April Pawluk, Joseph Bondy-Denomy, Vivian H. W. Cheung, Karen L. Maxwell, and Alan R. Davidson. A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. *mBio*, 5(2):e00896–14, May 2014.

[36] Pere Puigbò, Kira S. Makarova, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin. Reconstruction of the evolution of microbial defense systems. *BMC Evolutionary Biology*, 17:94, 2017.

[37] Benjamin J. Rauch, Melanie R. Silvis, Judd F. Hultquist, Christopher S. Waters, Michael J. McGregor, Nevan J. Krogan, and Joseph Bondy-Denomy. Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, 168(1–2):150–158.e10, January 2017.

[38] Rotem Sorek, C. Martin Lawrence, and Blake Wiedenheft. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annual Review of Biochemistry*, 82(1):237–266, 2013.

[39] Raymond H. J. Staals, Simon A. Jackson, Ambarish Biswas, Stan J. J. Brouns, Chris M. Brown, and Peter C. Fineran. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nature Communications*, 7:12853, October 2016.

[40] Aris-Edda Stachler and Anita Marchfelder. Gene Repression in Haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats) - Cas I-B system. *Journal of Biological Chemistry*, page jbc.M116.724062, May 2016.

[41] Aris-Edda Stachler, Israela Turgeman-Grott, Ella Shtifman-Segal, Thorsten Allers, Anita Marchfelder, and Uri Gophna. High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. *Nucleic Acids Research*, March 2017.

[42] Adi Stern, Leeat Keren, Omri Wurtzel, Gil Amitai, and Rotem Sorek. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genetics*, 26(8):335–340, August 2010.

[43] Daan C. Swarts, Cas Mosterd, Mark W. J. van Passel, and Stan J. J. Brouns. CRISPR Interference Directs Strand Specific Spacer Acquisition. *PLoS ONE*, 7(4):e35888, April 2012.

[44] Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14):6614–6624, August 2016.

[45] Marie Touchon, Sophie Charpentier, Olivier Clermont, Eduardo P. C. Rocha, Erick Denamur, and Catherine Branger. CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. *Journal of Bacteriology*, 193(10):2460–2467, May 2011.

[46] Marie Touchon and Eduardo P. C. Rocha. The Small, Slow and Specialized CRISPR and Anti-CRISPR of Escherichia and Salmonella. *PLoS ONE*, 5(6):e11126, June 2010.

[47] Yunzhou Wei, Rebecca M. Terns, and Michael P. Terns. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes & Development*, 29(4):356–361, February 2015.

[48] Ariel D. Weinberger, Christine L. Sun, Mateusz M. Pluciński, Vincent J. Denef, Brian C. Thomas, Philippe Horvath, Rodolphe Barrangou, Michael S. Gilmore, Wayne M. Getz, and Jillian F. Banfield. Persisting Viral Sequences Shape Microbial CRISPR-based Immunity. *PLoS Comput Biol*, 8(4):e1002475, April 2012.

[49] Ariel D. Weinberger, Yuri I. Wolf, Alexander E. Lobkovsky, Michael S. Gilmore, and Eugene V. Koonin. Viral Diversity Threshold for Adaptive Immunity in Prokaryotes. *mBio*, 3(6):e00456–12, December 2012.

[50] Edze R. Westra, Stineke van Houte, Sam Oyesiku-Blakemore, Ben Makin, Jenny M. Broniewski, Alex Best, Joseph Bondy-Denomy, Alan Davidson, Mike Boots, and Angus Buckling. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology*, 25(8):1043–1049, April 2015.

[51] Ido Yosef, Moran G. Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*, page gks216, March 2012.

[52] Haiyan Zeng, Jumei Zhang, Chensi Li, Tengfei Xie, Na Ling, Qingping Wu, and Yingwang Ye. The driving force of prophages and CRISPR-Cas system in the evolution of Cronobacter sakazakii. *Scientific Reports*, 7:40206, January 2017.