Is having more than one CRISPR array adaptive?

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

June 8, 2017

Abstract

Prokaryotes are ubiquitous across environments able to support life, and so are the viruses that infect them. Bacteria and archaea possess a variety of immune systems in order to defend themselves against these viral pathogens. One example is the CRISPR adaptive immune system, which is found across diverse prokaryotic lineages. Many prokaryotes have a CRISPR locus, and, surprisingly, many have more than one CRISPR locus. Here we examine how the multiplicity of CRISPR immune systems in a genome is related to the pathogenic environment. We use a comparative genomics approach to demonstrate that having more than one CRISPR array is adaptive on average across prokaryotes. This adaptive signature appears to be a function of the diversity of CRISPR arrays rather than their multiplicity alone. We then develop a simple deterministic model of CRISPR immune memory turnover. We show how a tradeoff between memory span and learning speed can lead to an optimal two-array solution in certain pathogenic environments.

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 [31, 30, 17]. This apparent immune redundancy, wherein individuals possess multiple different types of immune mechanisms or multiple instances of the same mechanism, is somewhat counterintuitive. Why have more than one immune system [19]? More specifically, why have more than one of the same type of immune system? Here we endeavor to answer that 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 [27, 14]. 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 [35, 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, 33, 32]. Thus the Cas proteins act as the machinery of the immune system, with specific proteins implicated in memory acquisition, crRNA processing, or immune targeting, and the CRISPR array can be thought of as the location in which memories are recorded.

CRISPR systems appear to be widespread across diverse bacteria 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 [28, 40, 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, 28]. 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 [18, 51]). Currently, the ecological factors shaping the distribution of CRISPR systems across environments and taxa are poorly understood.

One open question is whether or not the possession of multiple CRISPR systems by a single bacterial strain is adaptive, and if so how. Many bacteria have multiple CRISPR arrays, and some have multple sets of *cas* genes as well (e.g., [16, 10]). CRISPR and other immune systems are horizontally transferred at a high rate relative to other genes in bacteria [38], 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 there was an adaptive advantage to possessing multiple CRISPR systems, given that the phenomenon is so common. Additionally, in some groups a multi-CRISPR state appeared to be conserved over evolutionary time (e.g. [6, 1]). This is despite a deletion bias in microbial genomes [34, 23] that we would expect to remove extraneous systems over time. Here we provide the first large-scale evidence that bacteria and archaea tend to have more than one CRISPR array that is selectively maintained, based on publicly available genomic data. We then go on to compare several hypotheses for why having multiple arrays might be adaptive, using both comparative genomics and theoretical approaches. We propose 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`, [36]) 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 recomendations 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 [46]. 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

Consider the case where CRISPR arrays provide no selective advantage to a host but accumulate in a genome following a neutral process. If we assume that 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 independently and identically exponentially distributed with rate ν. This leads to an accumulation process of arrays in a genome that can be described as a simple linear birth-death process, which yields a Poisson stationary distribution of the number of arrays in the genome with rate $\lambda = \frac{\eta}{\nu}$. In reality, different individuals will experience different rates of horizontal transfer and loss due to different intrinsic (e.g. cell wall and membrane structure) and extrinsic factors (e.g. density of neighbors, environmental pH and temperature). While prokaryotic immune systems are gained and lost at a high rate in general, these rates vary largely across taxa [38]. Thus if we assume that the parameters determining array accumulation in a genome are generally constant over time but heterogeneous among genomes, then we can model the array dynamics within a genome i following the model described above with rate $\lambda_i = \frac{\eta_i}{\nu_i}$. The gamma distribution is often used to model variable rates, and is a flexible distribution with nice mathematical properties when applied to Poisson random variables. If we let arrays in a genome i accumulate following the process described above with rate $\lambda_i \sim \Gamma(\alpha, \beta)$, then the number of arrays X in any genome follows a negative binomial distribution $X \sim \mathrm{NB}(r, p)$ where $r = \alpha$ and $p = \frac{\beta}{1+\beta}$.

If we assume that in the absence of the *cas1* and *cas2* spacer acquisition machinery CRISPR arrays are non-functional and thus provide no selective ad-

vantage, then we can use the distribution of the number of CRISPR arrays in genomes lacking *cas1*, *cas2*, or both genes to estimate r and p. In the case where *cas1* and *cas2* are both present in a genome, we expect CRISPR arrays to confer an adaptive advantage. If we take the case where the possession of a single array is highly adaptive (i.e. viruses are present and will kill any susceptible host) but assume that additional arrays provide no additional advantage, then the array turnover dynamics after the addition of the first array will follow the immigration-death model described above. Thus the number of arrays in a given genome in the dataset should be $Y + 1$ where $Y \sim \mathrm{NB}(r, p)$. We can then estimate r and p by shifting the distribution of the number of CRISPR arrays in genomes possessing *cas1* and *cas2* so that the number of genomes with Y arrays is $f_Y = N_{Y+1}$ where the N_Y's are the actual observed counts.

In practice there are several ways to test our null hypothesis, that having a single functional array is adaptive but having more than one array is not. First, we can shift our with-*cas* distribution of array counts using the method described above and determine at what shift ($S^\star$) the mismatch between the empirical with-*cas* and *cas*-lacking array count distributions, measured as the sum of squared differences between the distributions, is minimized. Under our null hypothesis $S^\star = 1$, and a value of $S^\star > 1$ implies that having more than one array is adaptive.

We can also compare our parameter estimates for the Cas-lacking (N for "no cas") and single-shifted with-Cas (S for "shifted") distributions, assuming the negative binomial model described above. We would expect that $\hat{r}_N \approx \hat{r}_S$ and $\hat{p}_N \approx \hat{p}_S$ under our null hypothesis, but 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. This parameter-based 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.

Differential rates of 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 this possibility. 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 num-

ber 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 ($\overline{\Delta\mu_k}$) to detect if there is a significant difference from zero.

To validate our functional versus non-functional classification of CRISPR systems, we confirmed that CRISPR arrays in genomes with both *cas1* and *cas2* present tend to have more spacers, indicating a likely difference in spacer-uptake rate as we would expect if no-*cas* genomes cannot acquire spacers (S1 Fig, [13]). This difference in length is not as large as one might expect, possibly because some systems are able to acquire or duplicate spacers via homologous recombination [24] and arrays may have been inherited recently from strains with active *cas* machinery.

2.3 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):

$$\underbrace{\frac{dC_i}{dt}}_{\text{Spacers Targeting Phage } i} = \underbrace{a_i(t, C_i)}_{\text{Acquisition}} - \underbrace{\mu_L C_i \sum_j C_j}_{\text{Loss}} \tag{1}$$

where μ_L is the spacer loss rate parameter and a_i will be a function of time representing the viral environment. 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 [12, 15, 50]. 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 evironments, this indicates that having multiple-arrays may be the best solution in a given environment or set of environments that a bacterium is likely to encounter.

We analyze a simple case of two viral species where there is one "background" species representing the set of all viruses persisting over time in the environment:

$$\frac{dC_B}{dt} = \mu_A v_B - \mu_L C_B \left(C_F + C_B\right) \tag{2}$$

and another "fluctuating" species that leaves and returns to the environment after some interval of time:

$$\frac{dC_F}{dt} = \mu_A v_F f(t) - \mu_L C_F \left(C_F + C_B\right) \tag{3}$$

where μ_A and μ_L are the spacer acquisition and loss rates respectively, v_B and v_F are composite parameters describing the densities of each phage species in the environment multiplied by adsorption rate, and $f(t)$ is a binary function

that takes a value of one if phage F is present in the environment and zero otherwise.

We also can consider the phenomenon of priming in our model, wherein if an CRISPR system has a spacer targeting a particular viral species, the rate of spacer acquisition towards that species is increased [11, 45]. Thus

$$\frac{dC_B}{dt} = \mu_A v_B g(C_B) - \mu_L C_B \left(C_F + C_B\right) \tag{4}$$

and

$$\frac{dC_F}{dt} = \mu_A v_F f(t) g(C_F) - \mu_L C_F \left(C_F + C_B\right) \tag{5}$$

where

$$g(C_i) = \begin{cases} 1 & C_i < 1 \\ p & C_i \geq 1 \end{cases} \tag{6}$$

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 a CRISPR array (44%). Of these genomes, almost half have more than one CRISPR array (48%). When restricting ourselves only to genomes where the CRISPR spacer acquisition machinery was present (*cas1* and *cas2* present) 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 [29]). Only 2% of 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. 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 driven by just those few particular taxa. To control for this we randomly sub-sampled 10 genomes from taxa with greater than 10 genomes in the database. After sub-sampling, approximately 37% of genomes had more than one CRISPR array, and 65% of genomes with intact spacer acquisition machinery had more than one CRISPR array. Of those genomes with at least one array, 47% had more than one. 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 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. 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 overrrepresented taxa altered our parameter estimates, 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* [28, 29]) and one or fewer copies each of the genes necessary for spacer acquisition (*cas1* and *cas2*). Even when restricting our analyses to genomes with one or fewer sets of *cas* genes, it is clearly adaptive to have more than one (functional) CRISPR array, though the effect size is smaller in this case after subsampling ($\Delta\mu = 0.89 \pm 0.03$, 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 take a subset of our sub-sampled dataset restricted 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 ($\Delta\bar{\mu}_k = 0.44 \pm 0.14$). 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.

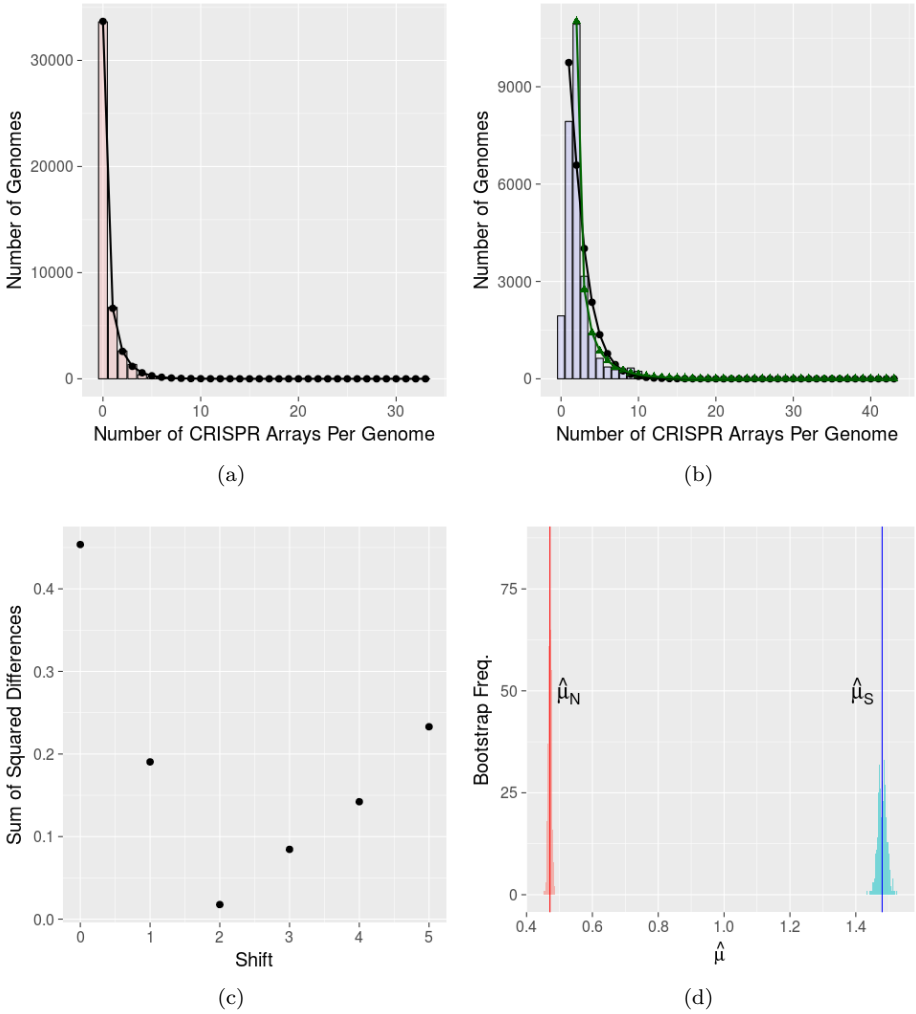


Figure 1: Having more than one CRISPR array is adaptive on average across prokaryotes. (a-b) Distribution of number of arrays per genome in (a) genomes that lacked *cas1*, *cas2*, or both, and (b) genomes that had *cas1* and *cas2* genes. In (a) black circles indicate the negative binomial fit to the 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.

			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	1.00	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.

3.3 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 peak in the distribution corresponding to dissimilar repeat sequences was significantly higher than in 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 when arrays share a set of *cas* genes. We did this by determining 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 also not driven by genomes with extremely low or high length-variable arrays (top and botton 5% excluded, $R^2 = 0.01041$, $p < 0.000698$).

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

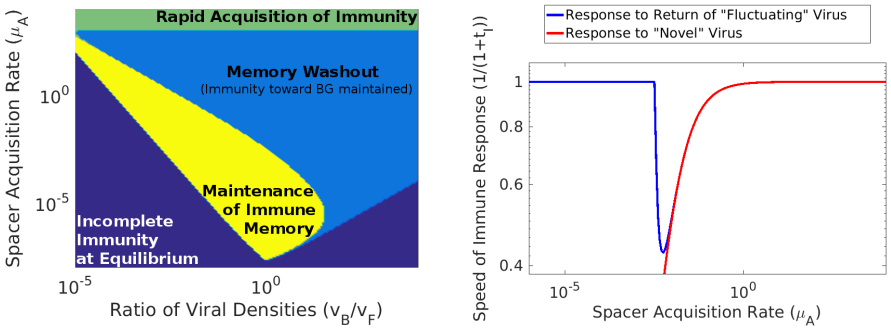
The evidence in Section 3.3 suggests that multi-array adaptiveness is linked to differences in consensus repeat sequences between arrays and that these differences may be associated with the spacer acquisition rate of each array. We hypothesized that having multiple systems with different acquisition 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 (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 acquisiton 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).

3.5 Taxon-specific signatures of adaptiveness

Several taxa in the dataset were represented by a sufficiently large number of genomes (> 1000) that varied in the presence of both *cas* genes and CRISPR-array counts that we were able to reliably perform our test for adaptiveness on each of these taxa individually. We found that among *Klebsiella pneumoniae* and *Staphylococcus aureus* genomes there was a signal of multi-system adaptiveness ($\Delta\mu = 0.60 \pm 0.06$, 0.63 ± 0.20 respectively), though relatively few of the *S. aureus* 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 ± 0.05 respectively), indicating that these species may occupy niches that favor single-array strains. *Salmonella enterica* had strongly negative $\Delta\mu$ values ($\Delta\mu = -1.05 \pm 0.11$), indicating that functional arrays are selected against in this taxon. Previous work has shown that CRISPR in *E. coli* and *S. enterica* appears to be non-functional as an immune system under natural conditions [48, 47]. All of these taxa are human pathogens, and can occupy a diverse set of environmental niches on the human body. It is unclear



(a) (b)

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 2.3, 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).

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). This is because the neutral array accumulation process cannot be estimated with confidence if most *cas*-lacking genomes are likely to have lost their *cas* machinery recently.

4 Discussion

4.1 Having multiple CRISPR arrays is adaptive across prokaryotic taxa

We show, for the first time, that, on average across prokaryotic taxa, having more than one CRISPR array adaptive. This general result holds true controlling for both overrepresented taxa and the influence of multiple sets of *cas* genes. It appears that this adaptiveness varies 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.

Our test for adaptiveness is based on the designation of arrays in genomes with both *cas1* and *cas2* genes present as "functional", and arrays in other genomes as "non-functional". This categorization is likely violated in some cases because (1) intact targeting machinery in the absence of acquisition machinery would still allow for preexisting spacers to confer immunity, (2) some CRISPR arrays may be conserved for non-immune purposes (e.g. [48, 26]), and (3) 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. Values for $S^\star$ roughly reflected the results for $\Delta\mu$, although they did not always detect weaker signals of adaptiveness (i.e., when $\Delta\mu < 1$), because we cannot assess the goodness-of-fit of partially-shifted distributions.

One potential phenomenon that could increase false positives in our test for adaptiveness is selection against having a CRISPR array in genomes lacking spacer acquisition machinery. This would violate our assumption of neutral accumulation and decrease $\hat{\mu}_N$. While there is a demonstrated deletion bias in prokaryotic genomes [34, 23], there is no reason we see that having a non-functional CRISPR array should be under strong negative selection because the associated costs should be low. 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. Despite this, the data appear to follow a negative binomial distribution quite well (Figs 1b and 1a, S2 Fig-S4 Fig).

4.2 Why have two CRISPR-Cas systems?

A prokaryote might gain an advantage from having multiple CRISPR systems either because (1) duplication of similar systems leads to improved immunity, or (2) having multiple systems with distinct features allows for the specialization of each system towards a specific type of threat. The relevance of different advantages depends on whether an individual has multiple sets of *cas* genes, CRISPR arrays, or both. We show that having multiple sets of *cas* genes is rare among prokaryotes, and that having multiple CRISPR arrays is adaptive regardless of the number of sets of *cas* genes, although this signal is particularly pronounced when multiple sets are present. Thus adaptive explanations that rely on multiple sets of *cas* genes can only be applied to a small number of taxa, and cannot explain the observed signature of adaptiveness in a large number of genomes.

In the case of the duplication of similar systems, immunity could be improved by an increased spacer acquisition rate, an increased rate of targeting, or a longer time to expected loss of immunity. In the case of an increased spacer acquisition rate, this effect would only be seen when multiple sets of *cas* acquisition machinery are present on a genome. Duplication of *cas* targeting genes could lead to more effective clearance of foreign genetic material from the cell via increased protein expression, but targeting has been shown to be very efficient in systems with only one set of targeting genes (e.g. [9]). Duplication of CRISPR arrays could lead to both an increased number of crRNA transcripts and a longer time to immune memory loss. In both cases array duplication will only confer an advantage if both arrays have spacers targeting the same viral species. Furthermore, the effectiveness of a crRNA may decrease in the presence of other competing crRNAs, meaning that multiple arrays could actually decrease targeting due to competitive interference between targets [42, 43].

Spacer loss in the CRISPR array most likely occurs via homologous recombination of repeat sequences [12, 15, 50]. Thus the time to immune loss will increase with the number of arrays targeting a particular viral species. Assuming that immunity towards a given virus in a single array has an exponentially distributed lifetime with expected value L (i.e., time to loss of all spacers targeting that virus in that array), in the absence of novel acquisitions the expected time to complete immune loss is $L\sum_{i=1}^{N}\frac{1}{i}$, where N is the number of arrays that initially target the virus in question. Clearly, the advantage conferred in terms of memory span decreases with each additional array, though this effect is important for the first few added arrays. In fact, it is more appropriate to model the lifetime of individual spacers with an exponential distribution such that the expected time to complete immune loss is $l\sum_{i=1}^{n}\frac{1}{i}$, where n is the total number of spacers in all arrays and l the expected lifetime of each spacer. Thus the relative advantage of multiple arrays is further reduced in the case where each array can have multiple spacers targeting the same virus, assuming that spacer

loss rates are similar across arrays (appropriate in the case of identical arrays near some equilibrium length). Additionally, we found a relationship between repeat diversity among arrays in a genome and the presence of *cas* acquisition machinery, possibly indicating a link between repeat diversity and multi-array adaptiveness. Such a diversity-driven effect would be inconsistent with the bet-hegding described above. It appears that CRISPR immune functionality is lost at a high rate in some prokaryotes [20], so that having multiple arrays could also represent a bet-heging strategy at the level of entire system-loss. That being said, this sort of bet-hedging also cannot explain the observed relationship between repeat diversity and *cas* presence.

Having multiple CRISPR systems might also be advantageous if having systems with different features is advantageous. For example viral proteins have been identified that target and inactivate the Cas targeting proteins of type I-E, I-F, and II-A systems [5, 37, 39]. By encoding multiple distinct sets of *cas* genes, hosts could evade the action of these anti-CRISPR proteins. Thus anti-CRISPR proteins have been proposed as a diversifying force in CRISPR system evolution and a possible explanation for system redundancy within strains [5]. However these anti-CRISPR proteins can often be extremely broadly acting, requiring surprisingly low levels of sequence identity (e.g., as low as 22% identity [37]) and sometimes even suppressing multiple system subtypes (e.g., I-E and I-F, [37]). Thus multiple *cas* gene sets will only be helpful if they are highly divergent within a strain, and potentially of different types with entirely different targeting genes.

Though only a small percentage of genomes had multiple *cas* signature genes, the majority of these genomes also had multiple types of such genes, consistent with a coevolutionary race between anti-CRISPR proteins and host in a small subset of strains. This is particularly surprising when contrasted with CRISPR arrays, since similar rather than different arrays tended to cluster within a genome, though this clustering was not seen to be adaptive. We also note that the inclusion of these multi-*cas* genomes in the dataset increased the effect size of our test for adaptiveness, despite their low relative representation in the dataset. Selection for multiple sets of *cas* genes will also select for multiple arrays, as arrays are generally *cas*-gene specific [22]. In any case, while coevolution with anti-CRISPR proteins remains an interesting candidate to explain why some prokaryotes have more than one CRISPR system, it cannot explain the signature for multi-array adaptiveness observed in the majority of the dataset.

It is reasonable to assume that as an array increases in length (i.e., the number of repeats increases) the rate of spacer loss will also increase because loss occurs via homologous recombination. A length-dependent spacer loss rate such as this would cause high acquisition rate systems to also have a high loss rate at equilibrium length. Thus increased uptake creates increased turnover of immunity as a side effect. In other words, there should be a tradeoff between the speed with which memory is acquired and the duration that a given memory lasts. Such an effect could lead to selection for both high activity (i.e., short term memory) and low activity (i.e., long-term memory) systems depending on the pathogenic environment that the host experiences (e.g., frequent viral extinction

and recurrence versus a steady background viral population). This tradeoff will disappear when the acquisition rate is high because memory becomes irrelevant in the limit of rapid immune acquisition. However, there are several reasons that the upper limit of immune acquisition rates should be constrained (e.g., limits on expression of *cas* genes and the CRISPR array, or autoimmunity [49, 21, 53, 25, 44], S2 Text). Even CRISPR arrays sharing a single set of *cas* genes may vary greatly in acquisition rate [41], meaning that a tradeoff hypothesis could explain the signature of adaptiveness in our multi-array single-*cas* dataset. Just as the data in our system suggests a link between consensus repeat and acquisition rate, differences in array length between arrays sharing a set of *cas* genes, but with slightly different repeats have been observed elsewhere [54].

Our mathematical model confirms that an acquisition rate versus memory span tradeoff can produce a bimodal landscape of optimal acquisition rates. This shows that, depending on the specific phage environment, having multiple systems optimized to solve either fast-learning or long-memory problems may be adaptive. The data indicate that there may be a link between array repeat diversity and multi-array adaptiveness, possibly mediated by relationship between consensus repeat sequence and spacer acquisition rates. This suggests that changes in repeat sequence have some functional role in CRISPR immunity, perhaps modulating spacer insertion rates. Mechanistically, it is unclear what would drive such a relationship. 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.

Many questions concerning CRISPR array multiplicity remain to be answered. Specifically, experimental verification that the consensus repeat sequence modulates spacer acquisiton rates is a first step towards validating the tradeoff mechanism we propose here. As more sequences and metagenomic datasets become available, it may be possible to explicitly link particular array configurations to specific features of the pathogenic environment or host lifestyle. Theoretical approaches that explore optimal immune system configurations will be useful in guiding researchers towards the appropriate data needed to compare the several hypotheses discussed here.

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 that so many CRISPR arrays would accumulate neutrally in a genome via horizontal transfer. We would expect that hightened rates of HGT should not be restricted to CRISPR arrays alone, so that 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 HGT alone cannot explain these outliers. It is possible that high rates of duplication of specific array types could lead to the observed pattern. Alternatively, there may be some adaptive advantage to array enrichment, though we are at a loss to what that might be.

Finally, our CRISPR-focused examination of immune system configuration could be expanded to include other types of prokaryotic defense, though progress has been made on this front by others (e.g. [18, 19, 21, 52]). While previous work has focused primarily on understanding why certain environments or lifestyles

favor certain immune strategies, or combinations of strategies, we emphasize that understanding how the multiplicity of immune systems evolves is largely an open question.

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] Kyle C. Cady, Joe Bondy-Denomy, Gary E. Heussler, Alan R. Davidson, and George A. O'Toole. The CRISPR/Cas Adaptive Immune System of Pseudomonas aeruginosa Mediates Resistance to Naturally Occurring and Engineered Phages. *Journal of Bacteriology*, 194(21):5728–5738, November 2012.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Victor Kunin, Rotem Sorek, and Philip Hugenholtz. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*, 8:R61, 2007.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.