The size of the immune repertoire of bacteria

Serena Bradde[a,b,1], Armita Nourmohammad[c,d,1], Sidhartha Goyal[e,1], and Vijay Balasubramanian[b,1]

[a]American Physical Society, Ridge, NY 11961; [b]David Rittenhouse Laboratories, University of Pennsylvania, Philadelphia, PA 19104; [c]Max Planck Research Group (MPRG): Statistical Physics of Evolving Systems, Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany; [d]Department of Physics, University of Washington, Seattle, WA 98195; and [e]Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada

Edited by Luca Peliti, Santa Marinella Research Institute, Roma, Italy, and accepted by Editorial Board Member Curtis G. Callan Jr January 23, 2020 (received for review March 1, 2019)

Some bacteria and archaea possess an immune system, based on the CRISPR-Cas mechanism, that confers adaptive immunity against viruses. In such species, individual prokaryotes maintain cassettes of viral DNA elements called spacers as a memory of past infections. Typically, the cassettes contain several dozen expressed spacers. Given that bacteria can have very large genomes and since having more spacers should confer a better memory, it is puzzling that so little genetic space would be devoted by prokaryotes to their adaptive immune systems. Here, assuming that CRISPR functions as a long-term memory-based defense against a diverse landscape of viral species, we identify a fundamental tradeoff between the amount of immune memory and effectiveness of response to a given threat. This tradeoff implies an optimal size for the prokaryotic immune repertoire in the observational range.

CRISPR-Cas | adaptive immunity | bacteria | phage | optimal memory

All living things from bacteria to the whale are under constant threat from viruses. To defend against these threats, many organisms have developed innate mechanisms that make it harder for infections to occur (e.g., by impeding entry of viruses through the cell membrane) or that respond universally to the presence of infections (e.g., through inflammatory responses) (1). While such innate defenses are very important, they likely cannot be as effective as defenses that specifically target particular infections. The challenge for developing specific responses is that they require a mechanism for learning to identify each incident virus and a mechanism for remembering the defended targets. Vertebrates implement such a learning-and-memory approach in their adaptive immune system, which produces novel antibodies through random genomic recombination, selects effective immune elements when they bind to invaders, and then maintains a memory pool to guard against future invasions. Recently, we have learned that some bacteria and archaea also enjoy adaptive immunity (2)—they maintain cassettes of spacers (snippets of DNA from previously encountered phages) and use these spacers via the CRISPR-Cas mechanism to identify and clear recurring infections by similar viruses.

The CRISPR-Cas mechanism for immunity has three stages (2, 3). Following a first encounter with a virus, after a successful defense through another mechanism or if the virus is ineffective for some reason (4), some of the Cas proteins recruit pieces of viral DNA and integrate these spacers into an array separated by palindromic repeated sequences in one of several CRISPR loci (5) of the bacterial genome. Each array defines a CRISPR cassette, and together, the cassettes carry a memory of past infections. In a second stage, CRISPR loci are transcribed as single operons through a mechanism that depends on the CRISPR type of the locus; the RNA strand is then cleaved into small interfering crRNAs (CRISPR RNAs) which form complexes with other Cas proteins (6). Finally, invading sequences are recognized by base-pairing with the CRISPR RNA. A successful match triggers cleavage of the viral genetic material.

In laboratory experiments that expose naive bacteria, which start without spacers, to carefully controlled environments, CRISPR cassettes are often small, consisting of a few spacers acquired during interactions with phages. Wild-type bacteria have larger cassettes which contain a few dozen to at most a few hundred spacers (2, 7–9). Metagenomic analysis of the human gut microbiome has revealed CRISPR cassettes with an average size of 12 spacers (10). A broader analysis of all sequenced bacteria and archeae found cassette lengths clustered in the 20 to 40 range (9). Similarly, a study of 124 strains of *Streptococcus thermophilus* revealed an average cassette size of 33 (11). RNA-seq screening of nine prokaryote species (12–14) showed significant RNA expression of dozens of spacers in each of several different CRISPR arrays in each species. There is a general decline in expression level of spacers with distance from the leader end of an array, but while promoter-proximal spacers are preferentially expressed, the decline away from the leader end is gradual, the expression pattern is sporadic, there can be internal promoters leading to enhanced expression of distal spacers, and sometimes there is even additional transcription in the reverse sense. Taking these patterns and the report of the common presence of multiple CRISPR arrays in a single genome (5) into account, typical prokaryotes enjoying CRISPR-based immunity show significant expression of a few dozen to a few hundred spacers.

These findings lead to a puzzle. Why do bacteria maintain such small memories in their adaptive immune systems given that they have probably been exposed over generations to thousands of species of phage (15)? Certainly, the size of the genome is not a constraint, since bacterial genome sizes lie in the range of millions of base pairs. Perhaps the organization of CRISPR immunity should be understood in terms of the dynamics of the coevolutionary chase in an extended encounter between an

Significance

Some bacteria possess an adaptive immune system that maintains a memory of past viral infections in the CRISPR loci of their genomes. This memory is used to mount targeted responses against later threats but is remarkably shallow: it remembers only a few dozen to a few hundred viruses. We present a statistical theory of CRISPR-based immunity that quantitatively predicts the depth of bacterial immune memory in terms of a tradeoff with fundamental constraints of the cellular biochemical machinery.

Author contributions: S.B., A.N., S.G., and V.B. designed research, performed research, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. L.P. is a guest editor invited by the Editorial Board.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: serena.bradde@gmail.com, armita@uw.edu, goyal@physics.utoronto.ca, or vijay@physics.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1903666117/-/DCSupplemental.

First published February 18, 2020.

evolving phage and its prokaryote host. In this framework, the authors of refs. 6, 8, and 16–19 find that 1) recently acquired spacers are most useful since the attacking phage is under selection for mutation of the viral protospacers targeted by CRISPR, 2) CRISPR is not useful if phages evolve too quickly and may even be lost if this immune mechanism incurs a fitness cost (6, 17, 19), and 3) prokaryotes need only carry five to seven spacers to optimally deploy a CRISPR-based defense against a coevolving phage (8, 16). Because of finding 3, given a cost for having more spacers either in fitness or due to the specific mechanisms of CRISPR-based immunity, prokaryotes should only carry about a half-dozen spacers (8, 16). Since typical prokaryotes express a few dozen to a few hundred spacers (5, 12–14), we need an alternative theoretical explanation for the size of CRISPR arrays.

A clue may be offered in the fact that older spacers often correspond to evolutionarily conserved regions in the genomes of persistent viruses (3, 11, 20–22). In fact, even if an older spacer is not a perfect match to a new invader, it may efficiently prime the acquisition of new immunity (23–30). These considerations suggest that CRISPR may be most useful as a long-term memory of a diverse viral landscape, for protection against reinfection by the original strains surviving in separate niches and for priming immunity against related strains, as opposed to being a mechanism for short-term learning of coevolving threats. With this assumption, we find that given limited resources for Cas complex formation (31), a deep memory imposes an opportunity cost by reducing the chance for the spacer specific to an invader to be activated in time to cleave an invading virus before it reproduces. Thus, there is a tradeoff between effectiveness of the defense and the depth of memory.

We analyze this tradeoff quantitatively and demonstrate that it predicts an optimal size for the total immune repertoire of a bacterium. This optimum is controlled by the number of Cas complexes that are available for cleaving viruses and by the diversity of the phage landscape. In the limit that the viral landscape is very diverse, the size of the immune repertoire is largely constrained by the number of Cas proteins that the bacterium can produce while carrying on its other functions (31). Using the biologically relevant range for Cas protein concentration in a bacterium, we show that the optimal number of expressed spacers should typically lie in the range of 10 to 100 spacers, consistent with genomic observations (7, 8, 9–14).

Results

CRISPR as a Probabilistic Memory of Phage. We consider a model of infection where bacteria encounter $j = 1, \ldots, K$ types of phage, each with probability f_j. For simplicity, all types of phage are taken to be equally infectious and to have similar growth rates, conditions that are easily relaxed. In the CRISPR mechanism for adaptive immunity, bacteria incorporate snippets of phage DNA (spacers) into a CRISPR cassette. Upon later infection, the bacteria recruit CRISPR-Cas complexes with spacers that match the invading phage to cleave the viral DNA (Fig. 1).

Suppose that the CRISPR cassettes contains L spacers in total and that an individual bacterium maintains a population of N_p Cas protein complexes that can be recruited to cleave invaders. The spacer configuration can be characterized in terms of a vector $\mathbf{s} = \{s_1 \cdots s_K\}$ with entries counting the number of spacers specific to each phage type. The total cassette size is the sum of s_j, i.e., $\sum_j s_j = L$, and quantifies the amount of immune memory stored by an individual bacterium. We describe the phage configuration in a given infection event as a vector $\mathbf{v}$ of length K with entries indicating presence (1) or absence (0) of each viral type. Finally, we define the configuration of complexes $\mathbf{d} = \{d_1 \cdots d_K\}$ as a vector with entries counting the number complexes specific to each phage during the CRISPR response. The total number

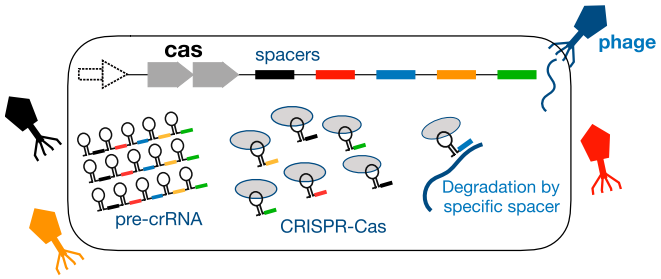

Fig. 1. CRISPR immunity in bacteria. A bacterium (bordered rectangle) with CRISPR machinery encounters a diverse set of phages (colors). The CRISPR-Cas locus is transcribed and then processed to bind Cas proteins (gray ovals) with distinct spacers (colors), thus producing CRISPR-Cas complexes. The complex with a spacer that is specific to the injected phage DNA (same color) can degrade the viral material and protect the bacterium from infection.

of complexes is the sum of d_j, i.e., $\sum_j d_j = N_p$. In terms of these variables, the probability of surviving a phage infection using the CRISPR-Cas defense mechanism is

$$P_{\text{survival}} = 1 - \left(\sum_{\mathbf{v}} p_V(\mathbf{v}) \sum_{s_1+s_2+\cdots=L} p_S(\mathbf{s} \mid L) \times \sum_{d_1+d_2+\cdots=N_p} [1 - \alpha(\mathbf{v}, \mathbf{d})] q(\mathbf{d} \mid \mathbf{s}) \right). \quad [1]$$

Here $p_V(\mathbf{v})$ is the probability of encountering the phage configuration $\mathbf{v}$, $p_S(\mathbf{s} \mid L)$ is the probability of having a cassette configuration $\mathbf{s}$ of length L, $\alpha(\mathbf{v}, \mathbf{d})$ is the probability of detecting all of the viral types present in $\mathbf{v}$ given the configuration $\mathbf{d}$ of complexes, and $q(\mathbf{d} \mid \mathbf{s})$ is the probability of producing the CRISPR-Cas configuration $\mathbf{d}$ given the set of spacers $\mathbf{s}$ and N_p Cas protein complexes.

Even if the number of phage types exceeds the number of complexes ($K \gg N_p$), bacteria can survive because we assume that a typical infection only involves a few viral types (or even just one). In this scenario, an infecting phage will attack a part of a large bacterial population. With cassettes sampling spacers randomly, at least some attacked individuals are likely to contain spacers specific to the phage and will thus survive. Innate mechanisms will also lead to survival of some bacteria without specific spacers, although we do not explicitly model this contribution to immunity. The surviving bacteria, and individuals which were not attacked, will replicate and maintain the population. In this context, when the next infection arrives, the cassettes in the bacterial population will again be effectively randomly drawn relative to the new infection. That is, although the cassettes will be enriched to reflect the previous infection, most of the spacers will still be randomly distributed so that cassettes in different individual bacteria will be largely uncorrelated. Over repeated encounters, the cycles of enrichment will lead to cassettes reflecting the distribution of phages. This scenario would be challenged if phages are very diverse ($K >> N_p$), new infections occur rapidly (faster than bacterial replication times of 1/2 to 1 h), and infections carry large viral loads (relative to the colony size). In this case the bacterial population will not have time to equilibrate between attacks and may need other defense mechanisms besides CRISPR to survive.

The form of the detection probability function $\alpha(\mathbf{v}, \mathbf{d})$ depends on the specific mechanism used by the CRISPR machinery to bind and degrade a phage. However, a critical number of specific complexes, d_c, is required for the CRISPR machinery to achieve targeting at the speeds measured in experiments (32).

This critical number depends implicitly on cell size, diffusion constants, timescales of phage infection and target recognition, and dissociation/association constants between spacers and protospacer sequences. Below this critical value, detection is less likely, and above the critical value the detection probability increases. We consider two functional forms for the probability that a particular phage type can be detected with d specific complexes: 1) a hard constraint $\alpha(d) = \theta(d - d_c)$, where $\theta(x)$ is step function and is 0 if $x < 0$ and 1 if $x > 0$, and 2) a soft constraint $\alpha(d) = \frac{d^h}{d^h + d_c^h}$. In both cases, d_c is an efficacy parameter that depends on biochemical rates and determines the threshold on d below which detection is rare and above which detection is common. The first functional form (step function) describes switch-like behavior where complexes bind to phage DNA if they exceed a certain concentration $d > d_c$. The second form mimics a Hill-like response, where the chance of binding increases gradually with the number of complexes. Here we allow the binding of Cas complexes with phage DNA binding to be cooperative. As the cooperativity h increases, the binding behavior becomes increasingly switch-like.

In our framework, we are assuming that spacer incorporation happens when the infecting phage is defective or if some other mechanism of immunity comes into play. We are then considering the subsequent probability of surviving phages due to CRISPR-Cas machinery when the spacer is already present. The probability of survival will increase when other forms of immunity (e.g., innate immunity and quorum-sensing effects) are also included. Importantly, CRISPR is not the first line of defense, and other antiphage mechanisms can precede or complement the CRISPR system (33). The net effect of these additional mechanisms can be modeled by including a nonzero baseline in the detection probability function α. Thus, our model describes the further survival advantage conferred by having CRISPR as a long-term memory of the phage landscape. Because of this, a bacterial population need not be in danger of extinction even if the CRISPR contribution to survival is relatively small.

There Is an Optimal Amount of Memory. In realistic settings we can make simplifying assumptions about the general model in Eq. **1**. For example, we can assume that successful infections of a bacterium by different phages occur with low probability and are independent. Of course, a given bacterium can be infected by multiple phages over its lifetime. Since the probability that a bacterium simultaneously encounters multiple phages is small, we assume that encounters are sequential (i.e., the viral configuration vector $\mathbf{v}$ has a single nonzero entry).

Second, we assume that a bacterium's lineage encounters many and diverse phage types over multiple generations, i.e., $K \gg 1$. Because phages mutate readily, there is subtlety about what defines a type. We use a functional definition—a type of phage is defined by its specific recognition by a given spacer. Sometimes, after a bacterium becomes immune to a phage, single point mutations in the virus can produces escapers that evade recognition. By our definition these escapers are effectively a new type of virus that the bacterial population must deal with sequentially in future infections (34–37).

Third, we assume that spacers are uniformly sampled from the phage distribution over time. In other words, we are assuming that prokaryotes pick up spacers from phages as they are encountered. So, if they encounter diverse viruses, they will have diverse arrays, and the phage distribution will be naturally reflected in the distribution of spacers. Incorporating such a distribution is straightforward, by assuming that the probability p_i that a spacer is incorporated is a function of the viral type i. However, this distribution is not known experimentally. Therefore, we make a minimal assumption that all phage types are equally likely and occur with probability $1/K$. This is a conservative assumption because bacterial immune memory confers the least advantage when faced with an unbiased (i.e., a minimally informative) phage environment. In effect, we focus on the long-term statistical features of immunity and not the the short-time coevolutionary arms race between bacterium and phage.

Finally, we assume that phage encounters from which spacers are acquired occur randomly. Thus, each spacer in the CRISPR cassette has a probability $1/K$ of being specific for a given phage. Since the cassette size is much smaller than the number of viral types ($L \ll K$), it also follows that the cassette will typically have one or no spacers that can target a particular phage type—in other words, $s_i = 0, 1$ (but see below for an analysis allowing multiple spacers from each phage).

In general, it is likely that the distribution of spacers is more uniform than the distribution of phages. Mechanistically, once a prokaryote has a spacer that works well for a given phage by targeting an evolutionarily conserved region in its genome, there will be less occasion in the long term to acquire many additional spacers from the same virus since the existing defenses work, although in the short term, CRISPR targeting may produce defective phages that encourage incorporation of additional spacers. (See *SI Appendix* for discussion of priming and the effects of multiple specific spacers.) On longer timescales, spacers from novel viruses are likely to be incorporated, leading to a distribution of spacers that is more uniform than the distribution of viruses. From a strategic standpoint, once there is a sufficiently effective defense against common threats, it is more statistically effective to devote the remaining resources preferentially to rare threats. Indeed, although phenomena like priming can lead to acquisition of multiple spacers against a given virus (23–30), several studies have shown that having just a few spacers from a given phage type is largely sufficient to neutralize reinfections (6, 8, 16–19). This means that the distribution of spacers should be more uniform than the distribution of pathogens—i.e., more weight should be given to rare infections than warranted by their frequency, again suggesting that a uniform distribution of spacers will be a reasonable approximation. Similar observations were made concerning the vertebrate adaptive immune system in refs. 38 and 39. RNA-seq screening in laboratory conditions has shown variable expression of spacers in CRISPR cassettes, typically accompanied by a gradual decline from the leader end, although there can be internal promoters leading to enhanced expression of distal spacers (12–14). We will approximate these sporadic expression patterns as a constant average across spacers whose expression is high enough to mount a defense against phages.

We derive an expression for the probability to survive a phage infection, given the cassette size L, the number of Cas complexes N_p, and the diversity of the phage population K (*Materials and Methods* and Eq. **3**). Across a wide range of parameters and for various choices of detection probability functions $\alpha(\mathbf{v}, \mathbf{d})$, we find that there is an optimal amount of memory, consisting of a few tens of spacers in the CRISPR cassette, to maximize survival probability (Fig. 2 and Fig. S1). The optimum occurs because there is a tradeoff between the amount of stored memory in CRISPR cassettes and the efficacy with which a bacterium can utilize its limited resources (i.e., Cas proteins) to turn memory into a functional response. If the CRISPR cassette is too small, bacteria do not remember past phage encounters well enough to defend against future infections. On the other hand, if the cassette is too large, Cas complexes bind too infrequently to the correct spacer to provide effective immunity against a particular invading virus. The optimal amount of immune memory (cassette size) should lie in between these two extremes, with details that depend on the phage diversity, the number of Cas complexes, and the detection probability function $\alpha(\mathbf{v}, \mathbf{d})$ (Fig. 2 and *Materials and Methods*).

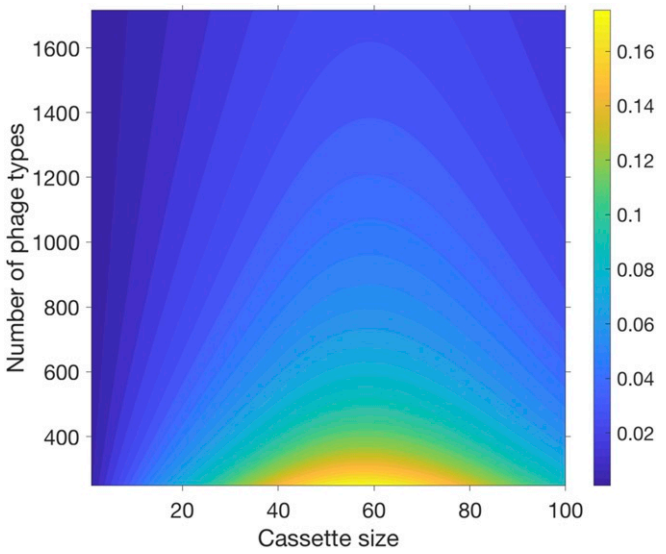

Fig. 2. There is an optimal amount of immune memory. The heat map shows the probability of surviving a phage infection, P_{survival}, as a function of the cassette size and phage diversity with $N_p = 700$ Cas complexes. P_{survival} can be interpreted as the fractional population size that will persist after sequential phage attacks if CRISPR is the only defense mechanism. The detection probability function is a step function $\alpha(d) = \theta(d - d_c)$ with threshold $d_c = 8$, implying that detection of a phage requires at least d_c complexes bound to the corresponding spacers. For any number of phage types, there is an optimal cassette size. See Fig. S1 for different choices of N_p, d_c, and functional forms for α.

Optimal Memory Depends on Phage Diversity. How does the optimal amount of CRISPR memory depend on the diversity of viral threats? If there are relatively few types of phage, an optimal strategy for a bacterium would be to maintain an effective memory for most threats and to match the viral variants with a cassette whose size grows with viral diversity. This matching strategy will eventually fail as the diversity of the phage population increases if the number of Cas proteins (N_p) is limited. We examined this tradeoff by measuring the optimal cassette size as we varied the viral diversity (K) while keeping the CRISPR machinery (N_p and the detection probability $\alpha(\mathbf{v}, \mathbf{d})$) fixed.

To characterize viral diversity, we defined a parameter $\kappa = K/N_p$ as the ratio of the number of phage types (K) and the number of Cas complexes (N_p). When phage diversity is low ($\kappa \leq 1$), the optimal amount of memory (number of spacers in a cell) increases sublinearly with viral heterogeneity (Fig. 3), approximately as a power law. When the detection probability $\alpha(\mathbf{v}, \mathbf{d})$ is nearly switch-like, the optimal cassette size scales approximately as $L \sim \sqrt{K}$ (Fig. 3 *A* and *B*; analytic derivation for $d_c = 1$ in *SI Appendix*). This implies that when viral diversity is low, the amount of memory should increase with the diversity, but it is actually beneficial not to retain a memory of all prior phage encounters. Forgetting some encounters will allow the bacterium to mount a stronger response against future threats by engaging a larger number of Cas complexes for the threats that are remembered. This sublinearity in the optimal amount of memory becomes stronger as the number of phage-specific CRISPR-Cas complexes necessary for an effective response, d_c, increases.

When phage diversity is high ($\kappa > 1$), the optimal amount of memory depends on the CRISPR mechanism via the response threshold d_c but is independent of viral heterogeneity so long as $d_c \geq 2$ (Fig. 3; see *SI Appendix* for discussion of the special case $d_c = 1$). In nature, phages are expected to be very diverse ($K \gg N_p$). Thus, our model predicts that the cassette size of a bacterium is determined by the expression level of Cas complexes N_p and the detection threshold d_c of the particular CRISPR mechanism that is used by the species.

Memory Increases with Detection Efficacy. Detection efficacy depends on two key parameters: 1) the detection threshold d_c and 2) the number of available Cas proteins N_p. In Fig. 4, we show that the optimal cassette size for defending against a diverse phage population ($\kappa = K/N_p \gg 1$) decays as a power law of the detection threshold $\sim (d_c/N_p)^{-\beta}$ with an exponent $\beta \simeq 1$ (*Materials and Methods*). This decay occurs because the trasncribed spacers compete to form complexes with Cas proteins; thus, having more distinct spacers effectively decreases the average number of complexes that would be specific to each infection. Thus, the smaller the cassette size, the more likely that the d_c specific complexes required for an effective CRISPR response will be produced. The optimal cassette size is a compromise between this drive toward having less memory and the drive to have a defense that spans the pathogenic landscape.

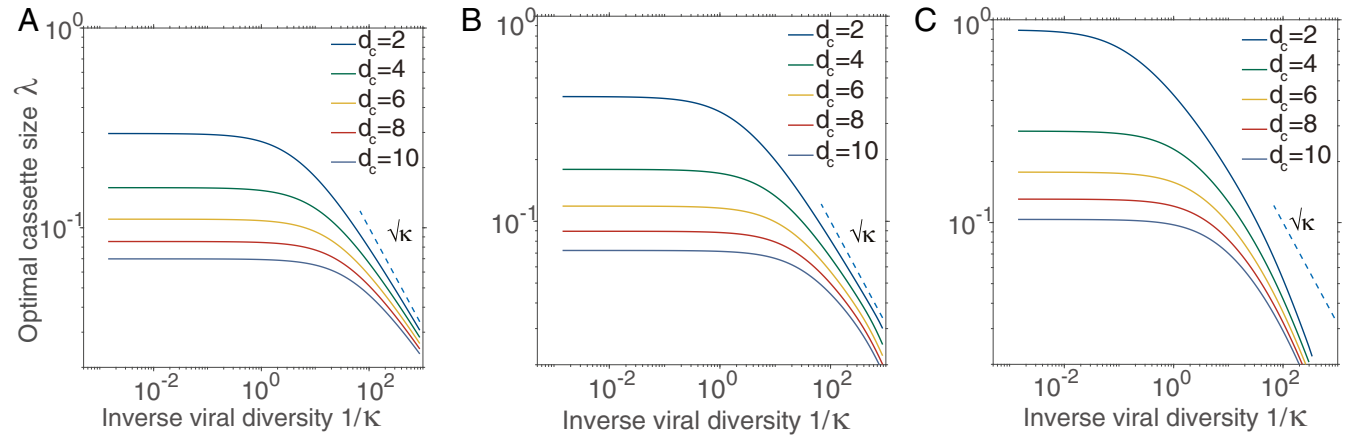

Fig. 3. Optimal amount of immune memory depends on the viral diversity. The panels show the optimal cassette size (L) relative to the number of Cas complexes (N_p) parameterized as $\lambda = L/N_p$, as a function of the viral diversity (K) relative to the number of complexes parameterized as $\kappa = K/N_p$. We examine CRISPR machineries with different detection probability functions: (*A*) switch-like detection probability with a step function $\alpha(d) = \theta(d - d_c)$ and a smoother model $\alpha(d) = d^h/(d^h + d_c^h)$ with (*B*) $h = 10$ leading to nearly switch-like detection probability and (*C*) $h = 2$ leading to a softer transition between low and high detection probability. Here d_c is an effective threshold on the number of complexes (d) for detecting phages with high probability.

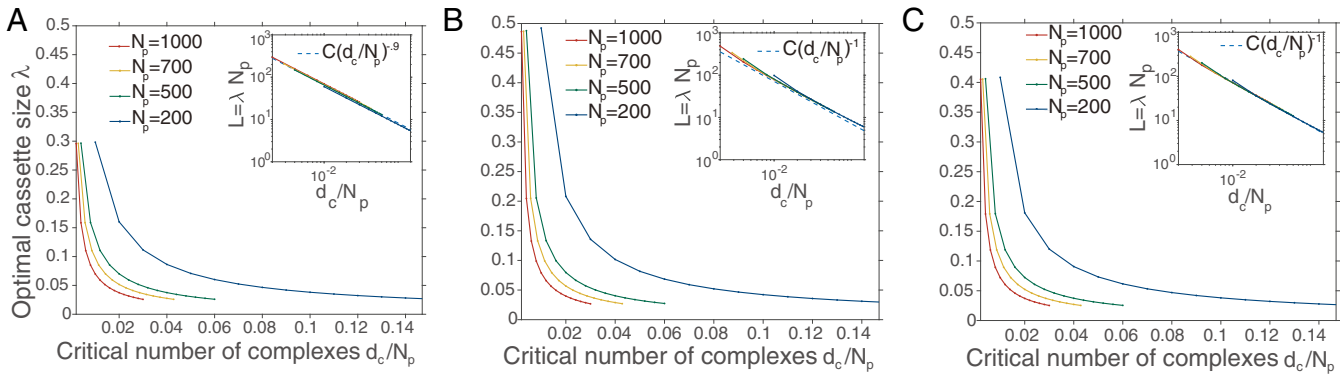

Fig. 4. Optimal amount of immune memory depends on the detection threshold. The figure shows the optimal cassette size relative to the number of Cas proteins, $\lambda = L/N_p$, as a function of the threshold to detect a phage, also relative to the number of Cas proteins, d_c/N_p. We consider different functional forms of the detection probability: (*A*) step function with a sharp detection threshold $\alpha(d) = \theta(d - d_c)$ and (*B* and *C*) Hill functions, $\alpha(d) = d^h/(d^h + d_c^h)$, with $h = 10$ in *B* and $h = 3$ in *C*. Phage types are taken to be 1,000-fold more numerous than the number of complexes ($\kappa = K/N_p = 1{,}000$). (*Insets*) Optimal cassette size scales as $L = C(d_c/N_p)^{-\beta}$ over a realistic range of values for the number of complexes and phage detection thresholds in a single bacterial cell. Best fits are shown in each case with (*A*) $\beta = 0.9$ and $C \sim 1$, (*B*) $\beta = 1$ and $C \simeq 0.7$, and (*C*) $\beta = 1$ and $C \simeq 0.8$.

If the detection probability depends sharply on the number of bound complexes with a threshold d_c (Fig. 4*A*), to a first approximation, fewer than d_c complexes bound to a specific spacer are useless, as detection remains unlikely, and larger than this number is a waste, as it would not improve detection. In this case, if the expression of the Cas protein was a deterministic process, it would be optimal to have a cassette with N_p/d_c spacers, each of which could be expressed and bind to exactly d_c complexes, predicting that $L = (d_c/N_p)^{-1}$. However, since gene expression is intrinsically stochastic, there would sometimes be more than d_c bound complexes for a given spacer and sometimes less. This stochastic spreading, arising partly due to finite size effects, weakens the dependence of the optimal cassette size on the threshold d_c, causing the exponent β and coefficient C to be slightly <1 in the optimal cassette size scaling $L \sim C(d_c/N_p)^{-\beta}$ (Fig. 4, *Insets*); see *Materials and Methods* for a more detailed derivation. If the detection threshold is soft, the CRISPR mechanism effectiveness is less dependent on having at least d_c complexes, specific to a phage. In addition, having a slightly higher number of complexes than the detection threshold can increase the detection probability. These effects combine to produce a scaling between the optimal cassette size and the detection efficacy, $L \sim (d_c/N_p)^{-1}$ (Fig. 4 *B* and *C*).

In summary, our model predicts that a more effective CRISPR mechanism (i.e., having lower detection threshold d_c or a larger number of complexes N_p) should be associated with a greater amount of immune memory.

Our model also provides an estimate for typical number of spacers per cell in bacterial populations countering a diverse set of pathogens (regime of $K \gg N_p$). Assume that the typical number of Cas complexes is $N_p \sim 1{,}000$, comparable to the copy number of other proteins in a bacterial cell (40, 41), and that rapid detection of an infecting phage requires a modest number of activated CRISPR-Cas complexes, with a detection threshold in range of $d_c \sim 10$ to 100 (32). Our model then predicts that the optimal immune repertoire should lie in the range of $L \sim 10$ to 100, consistent with empirical observations (2, 7–11).

Optimal Memory with Multiple Specific Spacers. CRISPR cassettes can contain more than one spacer specific to a given virus. This can happen, for example, due to priming, where the presence of some spacers that at least partially match an invading phage can lead to acquisition of additional spacers (23–30), increasing the effectiveness of the CRISPR-Cas system against recurrent or high-abundance viruses. On the other hand, having just a few spacers from a given phage type seems largely sufficient to neutralize reinfections (6, 8, 16–19), even from coevolving viruses. Furthermore, experimentally, wild-type CRISPR cassettes are known to target diverse phages rather than mostly having spacers targeting a few threats.

Thus, we generalize our model to allow $1 < s < s_{\max}$ spacers to be acquired from each phage type (details in *SI Appendix*). Following refs. 6, 8, and 16–19, the parameter $s_{\max}$ is expected to be a small number of the order of a handful. In this case, the optimal number of spacers remains similar to the result described above when there was at most one spacer for each phage (*SI Appendix*, Figs. S3 and S4). For completeness, we also tested the effects of allowing $s_{\max}$ to be unbounded. In this case, when phage species diversity is high as expected (15), the results are again unchanged. This is because having many spacers from one phage type comes at the cost of not detecting some other phage type when the number of complexes is limited by N_p. If phage diversity is sufficiently low, having many spacers for each virus does not exclude any unique phage type from being well represented; hence, the immune repertoire size in this scenario is not constrained.

Discussion

A bacterium's ability to neutralize phage attacks depends on the number of spacers stored in its CRISPR cassettes. In laboratory experiments, where viral diversity is limited, a few new acquired spacers per bacterium are enough to stabilize a bacterial population. A *S. thermophilus* population defending against continual phage attack acquired at most four new spacers over ~ 80 generations, with over 50% of the population having only one new spacer (42). A different experimental design where individual bacteria with different spacers were mixed found that an overall spacer diversity of ≥ 20 across the starting bacterial innoculum was enough to stabilize the population (43).

In natural populations, where viral diversity is high (15), spacer repertoires are significantly larger. CRISPR cassettes in the human gut microbiome contained 12 spacers on average (10), while a much broader analysis over all sequenced bacteria and archeae found cassette lengths in the 20 to 40 range (9). In agreement with this global analysis, 124 strains of *S. thermophilus* revealed an average cassette size of 33 (11). Meanwhile, RNA screening of diverse prokaryotes (12–14) showed expression of several dozen spacers residing in the multiple CRISPR cassettes present in most species enjoying this immune mechanism (5). All told, there is expression of a few dozen to a few hundred

spacers, but this is still just a fraction of the size of a typical bacterial genome.

One explanation for these observations may be that a bacterium in the wild encounters only a small set of phages (e.g., due to spatial constraints in habitats). Alternatively, it may be that the repertoires result from a functional tradeoff in the bacterial immune system. Our theory examines both how these considerations—phage diversity and tradeoffs in the CRISPR mechanism—affect the optimal amount of immune memory that a bacterium should store.

We identified two qualitatively distinct regimes for the statistics of optimal CRISPR immune memory. When phage diversity (K) is much lower than the number of available Cas proteins (N_p), the CRISPR cassette length L should increase sublinearly in K/N_p, approximately as the square root (Fig. 3). This is reminiscent of a previously identified optimal surveillance strategy of square-root biased sampling to catch rare harmful events while minimizing the wasteful resources spent on profiling innocents (38, 44).

When phage diversity is high ($K \gg N_p$), the number of available Cas proteins N_p limits effective use of memory to mount an immune response. In this regime, the optimal cassette size increases linearly with N_p, $L \sim N_p/d_c$. Here d_c quantifies the average number of Cas complexes specific to a given phage that is necessary to mount an effective defense. Why would the number of Cas proteins be limited in the first place? Perhaps there is in general a physiological cost to maintaining high levels of any protein within a cell. High sustained expression of the Cas system in particular may lead to autoimmune-like phenotypes, where a bacterium's lifespan is reduced because of the acquisition of spacers from its own genome (45). This may explain why, in some bacteria, the expression of Cas proteins is controlled by a quorum-sensing pathway that is triggered when the density of bacteria is high, making them more susceptible to infections (46–48).

To understand the tradeoffs that apply to CRISPR immune repertoires in the wild, we require a better understanding of both bacterial physiology and phage diversity in local environments. Quantifying phage diversity from metagenomic data is challenging, but there is evidence that phage species are very diverse (15). In particular, phage genomes are constantly changing as they undergo error-prone replications due to suboptimal use of hijacked bacterial replication machinery. Additionally, phages can swap pieces of DNA with their hosts as they are assembled and packaged inside an infected bacterium, leading to drastic changes in their gene synteny. These large-scale genomic changes challenge genome assembly and alignment techniques in quantifying phage diversity in a given community. At the same time, the expression levels of Cas proteins are not yet well quantified across broad bacterial families. However, these are topics of current research, and we can expect they will be addressed by ongoing advances in metagenomics and in high-throughput expression measurements in bacterial communities.

A causal test of our theory can also be conducted as follows. Create a library of cells where Cas expression is linked to different barcoded promotors (see, e.g., ref. 49) that allow expression level to vary from near knockout to several hundred copies. At the same time, integrate a known cassette with around 10 spacers in the genome (or on a plasmid). Create another plasmid library (mock viruses) where each plasmid has 1 spacer each, while the full population has all of the 10 spacers in near-equal numbers. These plasmids are designed such that if the cells' CRISPR system does not interfere, the cells are more likely to die (50). Then, the frequency of different survivor cells can be quantified via barcode sequencing after they are infected with the plasmid library. Every cell has the necessary spacers to clear any of the mock viruses, but according to our theory, the number of available Cas proteins will constrain effectiveness and thus the survival probability. An extension of this experiment could integrate cassettes of varying sizes and measure the optimal cassette size as a function of Cas expression level.

Several authors (6, 8, 16–19) have used detailed dynamical models to explore the possibility that CRISPR immunity is primarily useful as a short-term memory for defending against coevolving phages. In this context it may be enough to have just two spacers from an attacking phage to largely prevent escapers (18), although if phages mutate too quickly, the CRISPR mechanism seems unable to mount an adequate defense (6, 19). When phages mutate sufficiently slowly for CRISPR to be effective, only the most recently acquired spacers will be useful because the cocirculating phages will be under selection to mutate the regions of the genome targeted by the already acquired spacers (6, 8, 16–19). In view of this, ref. 16 restricted the CRISPR repertoire to consist of eight or fewer spacers, while ref. 8 assumed an exponential decay in expression of leader-distant spacers; both found that a repertoire of five to seven spacers is sufficient to optimize immunity. However, actual CRISPR repertoires consist of many more spacers (2, 7–9) distributed across multiple cassettes (5), with many dozens of spacers translated into RNA in a sporadic manner across each cassette despite a gradual decay in expression from the leader end (12–14). Thus, we pursued the alternative hypothesis that the primary role of CRISPR is to retain a long-term memory of previous invasions to guard against the diverse landscape of phage species (15). With this assumption, we arrived at an optimal size for the spacer repertoire in the measured range. The authors of ref. 5 similarly describe a dichotomy between long-term memory/slow learning and short-term memory/fast learning and suggest that CRISPR arrays with different spacer acquisition rates may partly play these different roles. If this is the case, it may be that the considerations of refs. 8 and 16 apply to some types of CRISPR arrays which deal with coevolving phages in conditions of low diversity, while our analysis applies to other arrays which maintain long-term memory useful in diverse viral landscapes with recurring infections.

If the CRISPR contribution to survival is too small, one could ask whether it is worth the effort to have this mechanism. Indeed, the authors of ref. 19 use a dynamic coevolution model to suggest that if phages evolve too quickly, CRISPR mechanisms should be eliminated if they have a cost. There is also some evidence that CRISPR may only be maintained in certain regimes of viral vs. bacterial density (51). From a theoretical perspective, the authors of ref. 52 describe scenarios in which CRISPR is not worth the cost. Our approach could be adapted in future work to analyze this question of when CRISPR is useful by including cost-benefit tradeoffs and interaction with ecological variables.

We do not know the diversity of the phage landscape faced by bacteria, but in a given environment a bacterium will only face a local pool of threats, and many of these might actually be targeting other species. After discounting for these factors, if there are ∼500 relevant phage types to defend against, CRISPR increases the probability of survival by 10% with a cassette of optimal size (Fig. 2), similar to the survival advantage found in refs. 8 and 16 for CRISPR defense against a single coevolving phage. Defending against ∼1,600 phage types, CRISPR with an optimal sized cassette confers a 3% advantage (Fig. 2). A p% difference in survival probability corresponds to a difference of approximately p% in the selection coefficient of a subtype with the better cassette during an evolutionary process. The standard replicator equations show that a subtype with a 1% selection benefit starting at 1% relative frequency will grow to make up 80% of the population within 600 generations, i.e., during just 10 to 12 d of *Escherichia coli* evolution (generation time of about 25 to 30 min). Indeed, selection strengths of just 1% are counted as

strong selection effects in population genetics. In fact, a subtype with a 3% selection benefit (the 1,600-phage scenario in Fig. 2) starting at 1% of the population will comprise 99% of the population within 600 generations. Thus, the increases in survival probability illustrated in Fig. 2 are likely to have substantial long-term effects on the population dynamics of bacteria equipped with CRISPR cassettes of the optimal size.

Vertebrates also possess an adaptive immune system that learns from past infections to defend against future threats. It has been suggested that pathogen detection in vertebrates is optimized by biasing the immune repertoire toward sensing rare infections with a higher chance than warranted by their frequency (38). This sort of optimization is unlikely to be relevant to individual bacteria given the small cassette size and high cost of CRISPR proteins. However, it is known that across a bacterial population spacer abundances are highly variable but distributed in a stereotypic way (53). It is possible that this stereotyped distribution represents an optimally adaptive immune strategy for the population as a whole. To study this question, our framework could be extended to analyze optimal strategies for distributed adaptive immunity in microbial communities.

For both vertebrates and bacteria, a major challenge is to characterize dynamics of the immune system as it chases a diverse pathogenic population that is itself evolving to evade detection by its hosts. This out-of-equilibrium process can last over an extended evolutionary period (54). Earlier work has attempted to account for such coevolutionary dynamics between prokaryotes with CRISPR and a few (typically one) species of phage (6, 8, 16–19). Here we pursued an alternative approach—we aimed to understand the statistical logic of a CRISPR-based immune system. In neuroscience, there is a similar distinction between computational models that address the dynamics of a neural network (e.g., the trajectory of neural population activity) and theoretical models that address the computational logic or goals of the same network (e.g., maximizing mutual information with the sensory input). Studying dynamics, as much as it can be insightful, requires a deeper experimental knowledge of the system, including the active components, how they interact, and the parameters of the dynamics, most of which are not well constrained by experiments. In the case of CRISPR-based immune systems, we still lack experimental data on the long-term coevolutionary parameters. Thus, it is important for the field to develop complementary modeling approaches which ask what the goal of the dynamics should be in order to provide effective immunity in the presence of general statistical and resource constraints. Thus, our approach aims to predict steady states of CRISPR immune system dynamics in a diverse environment. Evolution may be driven to select dynamics that achieve these steady states because this is what is useful for immunity, subject to tradeoffs and feedbacks associated with other system goals. Recently, a probability theory perspective of this kind has been applied to the logic of the adaptive immune repertoire of vertebrates (38, 39), but to our knowledge such an approach has not be applied to the study of CRISPR-based adaptive immunity.

Materials and Methods

Probability of Successful Immune Response. Applying the assumptions described in *Results*, the complete model (Eq. **1**) reduces to,

$$P_{\text{survival}} = 1 - \left(p_0 + \sum_d p_1 (1 - \alpha(d))\, q(d) \right) \quad [2]$$

where p_0 and p_1 are the probabilities for a bacterium to have zero or one spacer specific to the infecting phage, respectively; d is the number of specific complexes; $q(d)$ is the probability of producing d complexes; and $\alpha(d)$ is the probability that a cassette producing d specific complexes recognizes the phage. We assume that it is unlikely for a bacterium to carry more than one spacer against a given phage ($p_1 \approx 1 - p_0$) and take infections to be random events drawn from a pool of K distinct viruses. Hence, the probability that none of the spacers in a cassette of size L recognizes an invading phage is $p_0 = (1 - 1/K)^L \approx e^{-L/K}$. The survival probability is

$$P_{\text{survival}} = \left(1 - e^{-\frac{\lambda}{\kappa}}\right) \times \left(1 - \sum_{d<N_p} (1 - \alpha(d)) q(d)\right), \quad [3]$$

where $\kappa = K/N_p$ and $\lambda = L/N_p$ denote a normalized viral diversity and immune capacity, respectively. We assume that transcription events occur independently and are equally likely. Thus, $q(d)$ in Eq. **3** is given by a binomial density describing the probability of having d complexes specific to a given phage, given N_p Cas proteins and an equal probability of selecting any one of the L spacers to produce each complex. A successful detection typically requires activation of multiple complexes with a minimum (critical) number, d_c. Accordingly, we choose the detection probability function $\alpha(d)$ to be a threshold function that saturates to 1 at $d \geq d_c$ (*SI Appendix*).

Optimal Cassette Size. The optimal cassette size relative to the number of Cas proteins, $\lambda = L/N_p$, can be evaluated by optimizing the survival probability $\frac{\partial}{\partial \lambda} P_{\text{survival}}|_{\lambda^*} = 0$. In the biologically realistic regime, where the number of complexes is both large $N_p \gg 1$ and large compared to the cassette size $N_p/L \gg 1$, the binomial probability density for the number of specific complexes d in Eq. **3** can be approximated by a Gaussian $\mathcal{N}\left(\lambda^{-1}, \lambda^{-1}(1 - \frac{\lambda^{-1}}{N_p})\right) \simeq \mathcal{N}\left(\lambda^{-1}, \lambda^{-1}\right)$, up to quantities of order $\mathcal{O}(1/\sqrt{N_p})$. In this limit, the optimization criterion gives

$$0 = 1 - \lambda^* \sum_{d=0}^{d_c} \mathcal{N}\left(\frac{1}{\lambda^*}, \frac{1}{\lambda^*}\right)\left(\frac{3}{2\lambda^*} - \frac{d^2 - (\lambda^*)^{-2}}{2}\right), \quad [4]$$

where we assumed a diverse pool of viruses $\kappa \gg 1$ and a sharp recognition function $\alpha(d) = \theta(d - d_c)$. Approximating the sum in Eq. **3** with an integral with strict boundaries $[0, d_c]$, we arrive at an equation for the optimal cassette size λ^*,

$$1 = \frac{-e^{-\frac{1}{2\lambda^*}} + (1 + d_c\lambda^*) e^{-\frac{(1 - d_c\lambda^*)^2}{2\lambda^*}}}{2\sqrt{2\pi\lambda^*}} + \frac{1}{2}\left(\text{Erf}\left[\frac{1}{\sqrt{2\lambda^*}}\right] + \text{Erf}\left[\frac{d_c\lambda^* - 1}{\sqrt{2\lambda^*}}\right]\right). \quad [5]$$

In the limit that the cassette size is much smaller than the number of available complexes $L/N_p = \lambda \ll 1$, with $d_c\lambda$ finite, the optimal repertoire size scales inversely with the activation threshold $L^* = (1/2)\left(\frac{d_c}{N_p}\right)^{-1}$ in the Gaussian approximation of this section.

ACKNOWLEDGMENTS. V.B. is supported in part by a Simons Foundation grant in Mathematical Modeling for Living Systems (400425) for Adaptive Molecular Sensing in the Olfactory and Immune Systems, and by the NSF Center for the Physics of Biological Function (PHY-1734030). A.N. is supported by the Deutsche Forschungsgemeinschaft grant (SFB1310) for Predictability in Evolution, and MPRG funding through the Max Planck Society. S.G. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants Program (RGPIN 1505). V.B., S.B., and A.N. thank the Aspen Center for Physics, which is supported by NSF grant PHY-160761, for hospitality in the initial and final stages of this work.

1. S. J. Labrie, J. E. Samson, S. Moineau, Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).
2. R. Barrangou, L. A. Marraffini, CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54**, 234–244 (2014).
3. M. E. Bonomo, M. W. Deem, The physicist's guide to one of biotechnology's hottest new topics: CRISPR-Cas. *Phys. Biol.* **15**, 041002 (2018).
4. A. P. Hynes, M. Villion, S. Moineau, Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nat. Commun.* **5**, 4399 (2014).
5. J. L. Weissman, W. F. Fagan, P. L. Johnson, Selective maintenance of multiple CRISPR arrays across prokaryotes. *CRISPR J.* **1**, 405–413 (2018).
6. A. D. Weinberger *et al.*, Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput. Biol.* **8**, e1002475 (2012).
7. D. Rath, L. Amlinger, A. Rath, M. Lundgren, The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117**, 119–128 (2015).
8. A. Martynov, K. Severinov, I. Ispolatov, Optimal number of spacers in CRISPR arrays. *PLoS Comput. Biol.* **13**, e1005891 (2017).

9. I. Grissa, G. Vergnaud, C. Pourcel, The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinform.* **8**, 172 (2007).
10. T. C. Mangericao, Z. Peng, X. Zhang, Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls. *BMC Syst. Biol.* **10** (suppl. 1), 5 (2016).
11. P. Horvath *et al.*, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1401–1412 (2008).
12. D. L. Bernick, C. L. Cox, P. P. Dennis, T. M. Lowe, Comparative genomic and transcriptional analyses of CRISPR systems across the genus pyrobaculum. *Front. Microbiol.* **3**, 251 (2012).
13. H. Richter *et al.*, Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic Acids Res.* **40**, 9887–9896 (2012).
14. J. Zoephel, L. Randau, RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. *Biochem. Soc. Trans.* **41**, 1459–1463 (2013).
15. R. A. Edwards, F. Rohwer, Viral metagenomics. *Nat. Rev. Microbiol.* **3**, 504–510 (2005).
16. L. M. Childs, N. L. Held, M. J. Young, R. J. Whitaker, J. S. Weitz, Multiscale model of CRISPR-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin. *Evolution* **66**, 2015–2029 (2012).
17. B. R. Levin, Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet.* **6**, e1001171 (2010).
18. B. R. Levin, S. Moineau, M. Bushman, R. Barrangou, The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet.* **9**, e1003312 (2013).
19. J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, E. V. Koonin, Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *J. Bacteriol.* **195**, 3834–3844 (2013).
20. J. He, M. W. Deem, Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys. Rev. Lett.* **105**, 128102 (2010).
21. M. J. Lopez-Sanchez *et al.*, The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Mol. Microbiol.* **85**, 1057–1071 (2012).
22. C. L. Sun, B. C. Thomas, R. Barrangou, J. F. Banfield, Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J.* **10**, 858–870 (2016).
23. P. C. Fineran, E. Charpentier, Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. *Virology* **434**, 202–209 (2012).
24. R. Heler, L. A. Marraffini, D. Bikard, Adapting to new threats: The generation of memory by CRISPR-Cas immune systems. *Mol. Microbiol.* **93**, 1–9 (2014).
25. P. C. Fineran *et al.*, Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U.S.A* **111**, E1629–E1638 (2014).
26. D. C. Swarts, C. Mosterd, M. W. J. van Passel, S. J. J. Brouns, CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**, e35888 (2012).
27. E. V. Koonin, Y. I. Wolf, Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: Models and observations on virus-host coevolution. *Mol. Biosyst.* **11**, 20–27 (2015).
28. A. Ramachandran, S. Bailey, Memory upgrade: Insights into primed adaptation by CRISPR-Cas immune systems. *Mol. Cell* **64**, 641–642 (2016).
29. K. A. Datsenko *et al.*, Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).
30. E. Semenova *et al.*, Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7626–7631 (2016).
31. P. F. Vale *et al.*, Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. *Proc. Biol. Sci.* **282**, 20151270 (2015).
32. D. L. Jones *et al.*, Kinetics of dCas9 target search in Escherichia coli. *Science* **357**, 1420–1424 (2017).
33. S. Doron *et al.*, Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
34. H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390–1400 (2008).
35. J. F. Heidelberg, W. C. Nelson, T. Schoenfeld, D. Bhaya, Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **4**, e4169 (2009).
36. C. L. Sun *et al.*, Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol.* **15**, 463–470 (2013).
37. S. van Houte *et al.*, The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
38. A. Mayer, V. Balasubramanian, T. Mora, A. M. Walczak, How a well-adapted immune system is organized. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5950–5955 (2015).
39. A. Mayer, V. Balasubramanian, A. M. Walczak, T. Mora, How a well-adapting immune system remembers. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8815–8823 (2019).
40. R. Milo, What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **35**, 1050–1055 (2013).
41. M. N. Price, K. M. Wetmore, A. M. Deutschbauer, A. P. Arkin, A comparison of the costs and benefits of bacterial gene expression. *PLoS One* **11**, e0164314 (2016).
42. D. Paez-Espino *et al.*, CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. *mBio* **6**, e00262-15 (2015).
43. S. van Houte *et al.*, The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
44. W. H. Presse, Strong profiling is not mathematically optimal for discovering rare malfeasors. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1716–1719 (2009).
45. W. Jiang *et al.*, Dealing with the evolutionary downside of CRISPR immunity: Bacteria and beneficial plasmids. *PLoS Genet.* **9**, e1003844 (2013).
46. A. G. Patterson *et al.*, Quorum sensing controls adaptive immunity through the regulation of multiple CRISPR-cas systems. *Mol. Cell* **64**, 1102–1108 (2016).
47. N. M. Høyland-Kroghsbo, R. B. Mærkedahl, S. L. Svenningsen, A quorum-sensing-induced bacteriophage defense mechanism. *mBio* **4**, e00362-12 (2013).
48. N. M. Høyland-Kroghsbo *et al.*, Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 131–135 (2017).
49. J. B. Kinney, A. Murugan, C. G. Callan, E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9158–9163 (2010).
50. T. Wang *et al.*, Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nat. Commun.* **9**, 2475 (2018).
51. E. R. Westra *et al.*, Parasite exposure drives selective evolution of constitutive versus inducible defense. *Curr. Biol.* **25**, 1043–1049 (2015).
52. A. Mayer, T. Mora, O. Rivoire, A. M. Walczak, Diversity of immune strategies explained by adaptation to pathogen statistics. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8630–8635 (2016).
53. M. Bonsma-Fisher, D. Soutière, S. Goyal, How adaptive immunity constrains the composition and fate of large bacterial populations. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7462–E7468 (2018).
54. A. Nourmohammad, J. Otwinowski, J. B. Plotkin, Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genet.* **12**, e1006171 (2016).