PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Optimal number of spacers in CRISPR arrays

Alexander Martynov[1]*, **Konstantin Severinov**[1,2,3], **Iaroslav Ispolatov**[4]*

1 Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, Moscow, Russia, **2** Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States of America, **3** Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia, **4** Department of Physics, University of Santiago de Chile, Santiago, Chile

* jaros007@gmail.com (II); alexander.martynov@skolkovotech.ru (AM)

Abstract

Prokaryotic organisms survive under constant pressure of viruses. CRISPR-Cas system provides its prokaryotic host with an adaptive immune defense against viruses that have been previously encountered. It consists of two components: Cas-proteins that cleave the foreign DNA and CRISPR array that suits as a virus recognition key. CRISPR array consists of a series of spacers, short pieces of DNA that originate from and match the corresponding parts of viral DNA called protospacers. Here we estimate the number of spacers in a CRISPR array of a prokaryotic cell which maximizes its protection against a viral attack. The optimality follows from a competition between two trends: too few distinct spacers make host vulnerable to an attack by a virus with mutated corresponding protospacers, while an excessive variety of spacers dilutes the number of the CRISPR complexes armed with the most recent and thus most useful spacers. We first evaluate the optimal number of spacers in a simple scenario of an infection by a single viral species and later consider a more general case of multiple viral species. We find that depending on such parameters as the concentration of CRISPR-Cas interference complexes and its preference to arm with more recently acquired spacers, the rate of viral mutation, and the number of viral species, the predicted optimal number of spacers lies within a range that agrees with experimentally-observed values.

Author summary

CRISPR-Cas systems provide adaptive immunity defense in bacteria and archaea against viruses. They function by accumulating in prokaryotic genome an array of spacers, or fragments of virus DNA from previous attacks. By matching spacers to corresponding parts of viral DNA called protospacers, a CRISPR-Cas system identifies and destroys intruder DNA. Here we theoretically estimate the number of spacers that maximizes prokaryotic host cell survival. This optimum emerges from a competition between two trends: More spacers allow a prokaryotic cell to hedge against mutations in viral protospacers. However, the older spacers loose efficiency as corresponding protospacers mutate. For a limited pool of CRISPR-Cas molecular machines, keeping too many spacers leaves fewer of such machines armed with more efficient young (most recently acquired) spacers. We have shown that a higher efficiency of CRISPR-Cas system allows a prokaryotic

OPEN ACCESS

Citation: Martynov A, Severinov K, Ispolatov I (2017) Optimal number of spacers in CRISPR arrays. PLoS Comput Biol 13(12): e1005891. https://doi.org/10.1371/journal.pcbi.1005891

Editor: Claus O. Wilke, University of Texas at Austin, UNITED STATES

Received: June 1, 2017

Accepted: November 24, 2017

Published: December 18, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: II was supported by Fondo Nacional de Desarrollo Científico y Tecnológico (Chile) (http://www.conicyt.cl/fondecyt/) grant 1151524 and by SkolTech during his stay there. KS was supported by Skoltech-MIT Next Generation Program (https://skoltech.mit.edu/collaborative-projects/next-generation-program) and Russian Science Foundation grant №14-14-00988. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

cell to utilize more spacers, increasing the optimal number of spacers. On contrary, a higher viral mutation rate makes older spacers useless and favors shorter arrays. A higher diversity in viral species reduces the efficiency of CRISPR-Cas but does not necessary lead to longer arrays. Our study provides a new viewpoint at a variety of the observed array spacer number and could be used as a base for evolutionary models of host-phage coexistence.

Introduction

CRISPR-Cas systems provide prokaryotes with adaptive immunity against viruses and plasmids by targeting foreign nucleic acids [1–3]. Multiple CRISPR-Cas systems differing in molecular mechanisms of foreign nucleic acids destruction, *cas* genes, CRISPR repeats structure, and the lengths and numbers of spacers have been discovered [4, 5]. Yet the current understanding of diversity and function of CRISPR-Cas systems is far from being complete. The origins and, therefore, the targets of most spacers remain unknown [6–8]. The ubiquity of CRISPR-Cas systems in archaea compared to less than 50% presence in bacteria is also not well-explained [4, 9]. Evolutionary reasons for plethora of distinct CRISPR-Cas systems types, often coexisting in the same genome, remain largely unexplored [5, 10, 11]. It is also not clear why CRISPR arrays of some CRISPR-Cas systems contain only one or few spacers, while others have dozens or even hundreds of them [10–15]. It is commonly accepted that the number of spacers in an array is a result of a compromise between better protection offered against abundant, diverse, and faster evolving viruses by a larger spacer repertoire and a higher physiological cost of maintaining a longer array [16]. However, even the largest of the CRISPR systems contribute only 1% to the total size of a prokaryotic genome [11], so it is hard to imagine that adding or removing a few spacers would affect the growth rate in a noticeable way. Indeed, while there are various acknowledged sources of fitness cost for maintaining a CRISPR-Cas system [17, 18], none of them significantly depends on the number of the CRISPR spacers [11, 19, 20].

Virtually all models of prokaryotic and viral coevolution driven by CRISPR immunity include some representation of the number of CRISPR spacers. In some models the array content is limited by a maximal number of spacers (see, for example, [21], where such number is 8), or the number of spacers is determined dynamically as a result of competition between spacer acquisition and loss (such as in [22, 23]). For a given set of environmental conditions, such as the abundance and variety of infecting viruses, the dynamic determination of the optimal number of spacers often manifests itself as dominance of prokaryotic subpopulation with such arrays. At the same time, the number of spacers plays a major role in determining the complexity of simulation because it is usually required to check all possible pairwise spacer-protospacer matches to determine the immune status of a pair of prokaryotic and viral strains.

In this study, we propose a somewhat different view at the optimality of the number of spacers in CRISPR array. In particular, we ask a question of a rather idealized nature: What would be the number of spacers that maximizes protection of a given individual prokaryotic cell (rather than, for example, the survival of a prokaryotic species) from viruses? We show that the number of CRISPR spacers is primarily limited by “dilution” of CRISPR effector complexes carrying most immune-active CRISPR RNA with recently acquired spacers that target viral protospacers which had the least time to mutate. Our analysis requires a more detailed look at the kinetics of binding of CRISPR effector (a complex of Cas proteins with an individual protective CRISPR RNA, crRNA) to viral targets and distribution of crRNAs with particular spacers among the effectors. Since the origin and utility of the majority of spacers in each array are unknown, we made a simplifying assumption that all spacers in an array come from viral DNA and are used to repel viral infections. As another simplification instead of focusing on the actual evolution that occurs in ever-changing natural viral and prokaryotic communities, we compare the performance of arrays in their steady state for a given set of environmental parameters. We find that there exists a non-trivial optimal number of spacers, which maximizes the prokaryotic cell survival chances.

Models

Basic assumptions

Consider a prokaryotic cell with an active CRISPR-Cas system in a medium where phages capable of infection are present. The cell is attacked by individual viruses in a random and independent way: an attack is either repelled or kills the cell on a much shorter timescale than a typical time interval between subsequent attacks (Fig 1). We assume that CRISPR-Cas immunity is the only protection available against the infection and each infection which overcomes the CRISPR defense results in cell death.

The CRISPR array consists of a number of spacers acquired during previous viral attacks that did not result in the cell death and does not change over the timescale of analysis. Each spacer corresponds to a protospacer in DNA of viruses capable of infection. A match between a spacer and a protospacer is a necessary (but not sufficient) condition for efficient defense from infection. Protospacers may mutate, making now partially complementary spacer ineffective. Thus, it could be beneficial for a cell to pick up more than one spacer from each virus thus reducing the probability of failure of CRISPR-Cas system to recognize viral DNA [16].

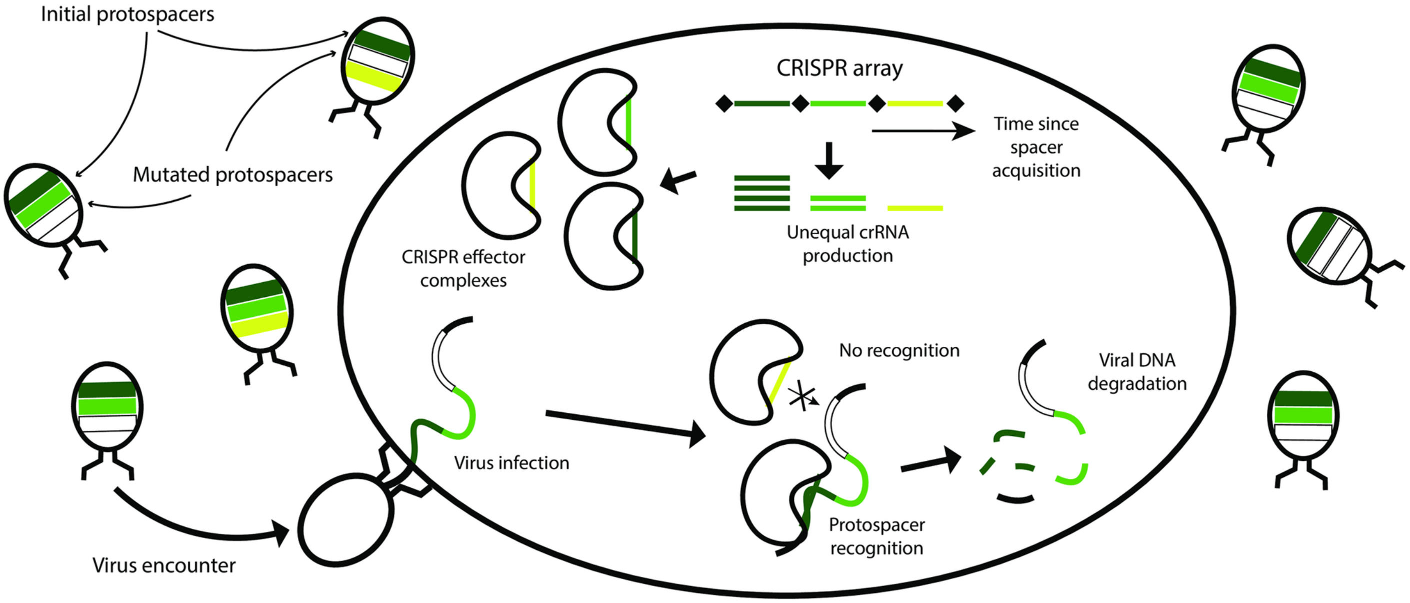

Fig 1. Functioning of CRISPR-Cas system. Three spacers are colored according to their age from the time of their acquisition, from dark green marking the youngest (the most recently acquired) spacer to yellow marking the oldest one (which was acquired the earliest). Phages carry protospacers colored similarly to their matching spacers; mutated protospacers are colored white. There are more mutated protospacers among protospacers matching older spacers than among protospacers matching younger ones. Inside the cell, bean-shaped objects are CRISPR effector complexes armed with individual crRNAs. Complexes with crRNA of younger spacers are more abundant than those with older ones. Viral DNA is shown to be simultaneously assessed by two CRISPR effector complexes: the dark green CRISPR spacer matches the non-mutated corresponding protospacer while the protospacer corresponding to the yellow spacer has mutated. The former interaction results in destruction of viral DNA while the latter leaves it intact.

https://doi.org/10.1371/journal.pcbi.1005891.g001

This allows the cell to hedge against mutation in single protospacer leading to more reliable recognition of the virus and increased probability of survival. It is intuitively appealing to arm more CRISPR effectors with newer, more recently acquired spacers rather than with the older ones so that the corresponding protospacers would have had less time to mutate. The older the spacer, the higher is the probability that the next encountered virus will have a corresponding protospacer mutated, leading to cell death. Indeed, there is a strong preference for spacers acquisition at one end of CRISPR array [24, 25]. As a result, spacers in natural arrays are ordered according to their age, with more recently acquired spacers located closer to promoter from which the array is transcribed. While the abundance of individual crRNAs is a complex function of their processing rate from pre-crRNA CRISPR-array transcripts and stability, promoter-proximal crRNAs are expected to be generally more abundant that promoter-distal ones [26]. This effect is expected from transcription polarity and made more pronounced by the palindromic nature of CRISPR repeats, which should promote transcription termination by RNA polymerase. Thus comes the second element of selective pressure over the number of CRISPR spacers: A too long array will "dilute" the concentrations of CRISPR effector complexes armed with crRNA of youngest (most recently acquired) and thus most efficient spacers, replacing them with crRNA of older spacers whose target protospacers had a longer time to accumulate mutations and thus become ineffective. For simplicity, we assume that a single mismatch between a spacer and its protospacer makes the corresponding crRNA completely ineffective in immunity [3]. While the reality is more complex and certain mutations in protospacers do not preclude recognition by the appropriately charged effector [27], mutations in protospacer adjacent motif [28, 29] or seed region [27] indeed abolish CRISPR interference and it is mutations of this kind that we consider in our work.

The optimal number of spacers may be thought of as emerging from competition between the opposing "more reliable recognition" and "dilution" trends. We ignore the fitness cost of maintaining a CRISPR array, often considered to be consisting of two parts: spacer-number-independent and spacer-number-dependent [21, 22]. While duplication of CRISPR-Cas system DNA must have its cost, yet every new spacer constitutes a very small part of CRISPR-Cas DNA (which itself is a small part of cellular genome) and such cost is ignored.

To summarize, we try to determine the optimal number of spacers in a CRISPR system illustrated in Fig 1 under the following simplifying assumptions:

- The cutting of viral DNA is possible when there is a perfect match between the spacer and protospacer, and a single mismatch makes the spacer-protospacer pair useless for cell protection/CRISPR interference [27–29].
- Probability for a CRISPR effector complex to contain crRNA with a particular spacer decreases exponentially with the age of the spacer.
- The total number of CRISPR complexes in a cell is constrained and independent of the number of spacers in an array. For simplicity, we further assumed that the number of CRISPR effector complexes is constant in time. There is evidence that *cas* genes expression in some systems is regulated in vivo depending on the external conditions [30, 31] and, in particular, may be increased during viral invasion [32, 33]. However, the constant Cas protein levels that we consider in the following could be viewed as maximal concentrations of these proteins in their "fully active" state or suitable time averages.
- A single encounter between CRISPR-effector and virus DNA resolves on a shorter timescale than the time between subsequent encounters.

- There is only a single copy of viral DNA inside the cell upon infection, i.e., the multiplicity of infections is low.
- We do not take into account any fitness costs of maintaining an array of a given spacer number [19, 20].
- The number of spacers in a CRISPR array does not change during the course of our thought experiment, i.e. on the timescale of several viral infections. For the single-virus case this does not imply that the array composition remains unchanged, it requires only that the number of spacers stays the same. For the multiple-virus case (see subsections "Analytical results: Multiple viral species" and "Numerical results: Multiple viral species") there is an additional assumption that the array composition does not change, i.e., there is no CRISPR adaptation on the timescale of several virus attacks. Given that the rate of naïve adaptation is very low [34] and that the primed adaptation is not considered in our main analysis and has only been described for several subtypes of Type I CRISPR-Cas systems, this assumption does not seem to be unreasonable and should apply to at least some CRISPR-Cas systems, particularly, Type II.

Probability of interference

Assume that a cell carries an array consisting of CRISPR spacers which we number in the direction of age such that the most recently acquired spacer is assigned number 1. The cell is being attacked by a virus and CRISPR defense comes into play. The probability B_i for CRISPR effector charged with crRNA with spacer i to bind to the corresponding protospacer (or the fractional occupancy of the protospacer) is controlled by competition between binding and dissociation events which are described by the first and second terms in the right-hand side of the following kinetic equation,

$$\frac{dB_i}{dt} = k^+(1 - B_i)C_i - k^- B_i. \tag{1}$$

Here k^+ and k^- are the association and dissociation rate constants for a matching spacer-protospacer pair and C_i is the copy number (uniquely related to its concentration since the volume of the cell is constant) of CRISPR effectors carrying the ith spacer crRNA. The steady state binding probability (or the fraction of time the corresponding protospacer is recognized by CRISPR effector) is

$$B_i = \frac{k^+ C_i}{k^+ C_i + k^-} = [1 + k^-/(k^+ C_i)]^{-1}. \tag{2}$$

For simplicity, we do not separately consider the transport phase of the spacer-protospacer binding, i.e. the time it takes a CRISPR effector and viral DNA to diffuse towards each other, and account for this phase by adjusting the k^+ and k^- constants. Now we compute how C CRISPR effectors present in the cell pick up crRNAs with particular spacers. We have postulated that the number of effector complexes that acquired spacer i decreases exponentially with increase of i. That is, each next spacer is δ times less likely to be present in CRISPR effector complex than its younger neighbor. We will further refer to δ as "crRNA decay coefficient" since we assume that the exponential decrease in the number of crRNA molecules with a defined spacer causes the corresponding decrease in the number of CRISPR effector complexes with this crRNA [26]. Hence the number of effector complexes C_i with crRNA with spacer i is

$$C_i = C_1\delta^{i-1}. \tag{3}$$

We determine C_1 from the condition that the total number of CRISPR effector complexes is C by summing the corresponding geometric progression

$$C_i = C\delta^{i-1}\frac{1-\delta}{1-\delta^S} \tag{4}$$

where S is the total number of spacers in the array.

Substituting (4) into (2) produces a complete expression for the binding probability between the ith spacer-protospacer pair,

$$B_i = \left(1+\frac{1}{\beta}\frac{1}{\delta^{i-1}}\frac{1-\delta^S}{1-\delta}\right)^{-1}. \tag{5}$$

Here $\beta \equiv Ck^+/(k^-)$ is the dimensionless coefficient which determines the "binding efficiency" of CRISPR effector. The larger β, the larger fraction of time the effector spends bound to matching protospacer. The biological meaning of β becomes clear if one considers a CRISPR array consisting of a single spacer. Then the binding probability becomes the function of β only,

$$B = \frac{1}{1+1/\beta}. \tag{6}$$

In such a case, the binding probability depends on how β compares to 1: If $\beta \gg 1$, the binding probability saturates to its maximum equal to 1, while if $\beta \ll 1$, the binding probability becomes proportional to β. For $\beta = 1$ the binding probability is precisely 1/2.

Assume that binding of every CRISPR effector to its matching protospacer proceeds independently of binding by other effectors to theirs, i.e., protospacers are well-separated in viral genomes. The total rate of interference is then proportional to the sum of binding probabilities of matching spacer-protospacer pairs, and the probability of survival of viral DNA $P(t)$ decays with a simple exponential kinetics,

$$\frac{dP(t)}{dt} = -aP(t)\sum_i B_i; \quad P(t) = \exp\left(-at\sum_i B_i\right). \tag{7}$$

Here a is the viral DNA degradation rate constant, which we consider to be a fixed property of a CRISPR-effector universal for all spacer-protospacer pairs. Hence the probability of successful interference is

$$I = 1 - P(\tau), \tag{8}$$

where τ is the effective time of interference, roughly equal to the time of the duplication of viral DNA. In other words, for successful termination of infection, the CRISPR effector complexes have to destroy the viral DNA before or during the first round of its duplication. Destruction of individual viral genomes at later times can not prevent the runaway viral DNA replication and productive infection. Introducing a dimensionless parameter $\chi \equiv \tau a$, which characterizes the interference efficiency, turns Eqs (8 and 5) into

$$I = 1 - \exp\left[-\chi \sum_i B_i\right] = 1 - \exp\left[-\chi \sum_i \left(1 + \frac{1}{\beta}\frac{1}{\delta^{i-1}}\frac{1-\delta^S}{1-\delta}\right)^{-1}\right]. \tag{9}$$

Survival probability

Assume that a virus infecting a cell at a given moment is drawn from a big pool with a probability of infection proportional to the concentration of its type v and that infections by different viruses are independent of each other. Then the probability A_k to experience k infections over time t is given by a Poisson distribution with the average number of infections rNt scaling linearly with time,

$$A_k(t) = \frac{(rNt)^k}{k!}\exp(-rNt), \tag{10}$$

where r is a proportionality coefficient considered to be the same for all viruses and N is concentration of the viral particles. To survive during a given time, each cell needs to repel all infections happening within this time, hence the probability of survival till time t is

$$\sum_{k=0}^{\infty} A_k(t) I^k = \exp[-rNt(1-I)]. \tag{11}$$

Here I, defined in Eq (9), is the probability to survive a single infection, i.e., the probability of successful CRISPR interference. From our assumption that viruses infect independently of each other it follows that the probability $E(t)$ for a cell to survive in the medium with several different viruses with concentrations v_j is given by the product of survival probability determined for each virus separately,

$$E(t) = \prod_j \exp[-rN_j t(1-I_j)]. \tag{12}$$

This is sketched in Fig 2. The probability of CRISPR interference with a single infection I_j is defined as in (9) with the sum running over all spacers taken from the jth virus. In the following we use $E(t)$ as the measure of overall CRISPR system performance.

Results

Analytical results: Single viral species

To illustrate and further develop the general statement (12), consider a scenario of a single viral species infecting a cell that has a CRISPR array with just two spacers. The immunity depends on the mutation status of corresponding protospacers in viral population. In this model, the mutation status of the spacer will be defined as the fraction of mutated protospacers in the viral population. We denote by m_1 and m_2 the probabilities for the first and second protospacers to remain mutation-free and thus recognizable by CRISPR effectors. If the total concentration of viral particles is N, the concentration of the "wild type" variant without any mutations is m_1m_2N, the concentration of the variant with mutation in the second protospacer is $m_1(1 - m_2)N$, the concentration of the variant with mutation in the first protospacer is

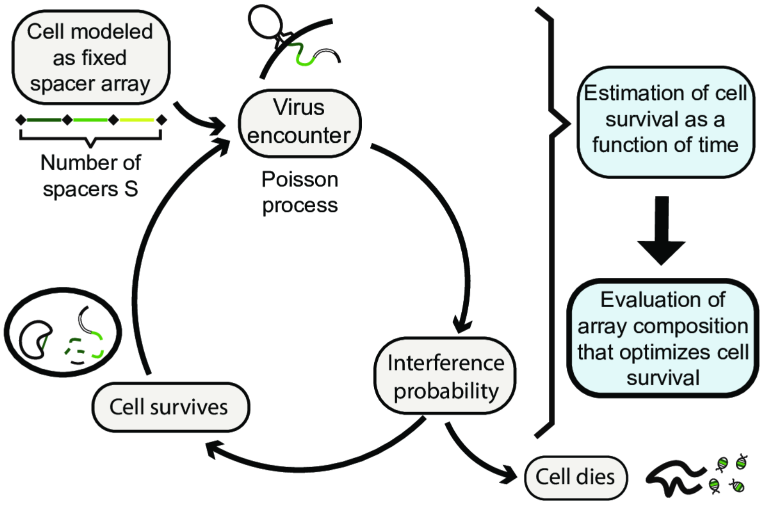
Cell modeled as fixed spacer array
Number of spacers S
Virus encounter
Poisson process
Interference probability
Cell survives
Cell dies
Estimation of cell survival as a function of time
Evaluation of array composition that optimizes cell survival

Fig 2. Scheme of calculations. A cell with $S = 3$ CRISPR spacers encounters viruses as a Poisson process with an average rate rN. During each encounter there is either a successful interference with probability I or the cell dies with probability $1 - I$. We evaluate the probability $E(t)$ of the cell to survive till time t as the measure of performance of its CRISPR-Cas system.

https://doi.org/10.1371/journal.pcbi.1005891.g002

$m_2(1 - m_1)N$, and the concentration of the variant with mutations in both protospacers, i.e., an escape mutant not subject to CRISPR interference, is $(1 - m_1)(1 - m_2)N$. From Eqs (9 and 12) and our assumption that a mutation in protospacer renders the corresponding spacer completely inefficient, it follows that the survival probability in such case is

$$\begin{aligned} E(t) = \exp(-rNt\{m_1 m_2 \exp[-\chi(B_1 + B_2)]+ \\ +m_1(1 - m_2)\exp[-\chi B_1] + m_2(1 - m_1)\exp[-\chi B_2] - (1 - m_1)(1 - m_2)\}). \end{aligned} \tag{13}$$

The last term in the exponent corresponds to the probability to experience no infection by viruses with both mutated protospacers (in which case $I_4 = 0$ since such an infection would result in cell death). Transforming the expression in the exponent, we obtain

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{2}\{1 - m_i[1 - \exp(-\chi B_i)]\}\right)\right]. \tag{14}$$

This expression has a simple probabilistic interpretation: The ith term in curly brackets describes the probability of failure of CRISPR effector complexes armed with the ith spacer crRNA. The product of such terms describes the probability of failure of all CRISPR effectors and thus the death of the cell. The expression (14) is the probability for the Poisson process of "failures" of CRISPR system to have zero counts or no failures at all, which translates into survival of the cell. Mutual independence of encounters with different mutation variants of the virus simplifies the survival probability of the cell to the probability of not to be affected by the "average" encounter repeated rNt times. This simple interpretation allows us to generalize (14) to cases of arrays containing more than 2 spacers, replacing the upper limit of the product by an actual number of CRISPR spacers S,

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{S}\{1 - m_i[1 - \exp(-\chi B_i)]\}\right)\right]. \tag{15}$$

The Eqs (12) and (15) are universal and are applicable to a variety of scenarios involving CRISPR immunity. For example, (12 and 15) can serve as a base for evolutionary dynamics models, where the mutation status of protospacers and the composition of CRISPR array are determined dynamically for each viral and host strain. In addition to their more traditional population dynamics applications, such models can mimic the evolution of various parameters of CRISPR systems and even more intricate features like the preference to acquire spacers from particular parts of viral genomes [35] or the co-evolution of CRISPR individual immunity and altruistic abortive infection mechanisms [36]. However, it is hard to visualize the conclusions that follow from (12 and 15) in their general form due to the large number of generally unknown parameters m_i.

To reduce the number of independent parameters in Eq (15) and in the following expressions for the survival probability, we estimate m_i. We assume that spacers were acquired to the array in a periodic fashion, that is, the time intervals t_{ins} between subsequent acquisition of spacers were the same. The probability for a protospacer to remain mutation-free decreases exponentially with time, and the "age" of the ith protospacer is proportional to i. Hence, the probability of a perfect match for the ith spacer-protospacer pair at the middle of the time interval between spacer acquisitions can be approximated as $\mu^{i-1/2}$. Here $0 < \mu < 1$ is the probability for a protospacer in viral DNA not to undergo any mutations during t_{ins} and −1/2 in the exponent stands for assessing the cell survival probability in $t_{ins}/2$ time units after the acquisition of the last spacer, i.e. in the middle of the interval between spacer acquisitions. The parameter μ depends on genetic and environmental factors such as the rate of mutations in viral DNA, the size of the viral population, the size of protospacer, and the average rate at which cells acquire new spacers. Eq (16),

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{S}\{1-\mu^{i-1/2}[1-\exp(-\chi B_i)]\}\right)\right], \tag{16}$$

together with the binding probability (5), completely define the survival probability of a cell with a given number of spacers S as a function of dimensionless parameters μ, χ, δ and β. Note that the optimal number of spacers does not depend on the total time of observation t that was used for cell survival evaluation: In Eq (16) the position of the maximum of $E(t)$ is determined by the maximum of the product in the exponent and is independent of rNt.

Numerical results: Single viral species

A typical dependence of survival probability $E(t)$ on the crRNA decay coefficient δ and the number of spacers S is shown in Fig 3. For this example, we inferred the interference probability $I_1 \approx 0.5$ of a single spacer array from the experimental data [35] (see S2 Appendix for details). While the exact values of binding efficiency β and interference efficiency χ cannot be determined separately from I, we set them to some intermediate values $\beta = 1$ and the $\chi = 1.4$ that reproduce the measured I_1. It is shown in [37] that the interference rate per DNA molecule noticeably drops when the copy number of DNA molecules increases from one to a few, which indicates a relative shortage of Cas effector complexes and supports our choice for an intermediate value of β. See S2 Appendix for an example which uses a different pair of β and χ for the same I. The probability for a protospacer not to mutate over the typical period between spacer acquisition was chosen to be $\mu = 0.9$. The typical number of infections over the time of observation was $rNt = 5$. It follows from Fig 3 that the survival is maximized for crRNA decay coefficient $\delta \approx 0.7$ and the number of spacers $S = 6$. In panel B the dependence of $E(t)$ vs. S is shown for several values of δ. Curiously, for low δ, the survival $E(t)$ does not noticeably decrease for large S. It happens because of the exponential suppression in frequencies of

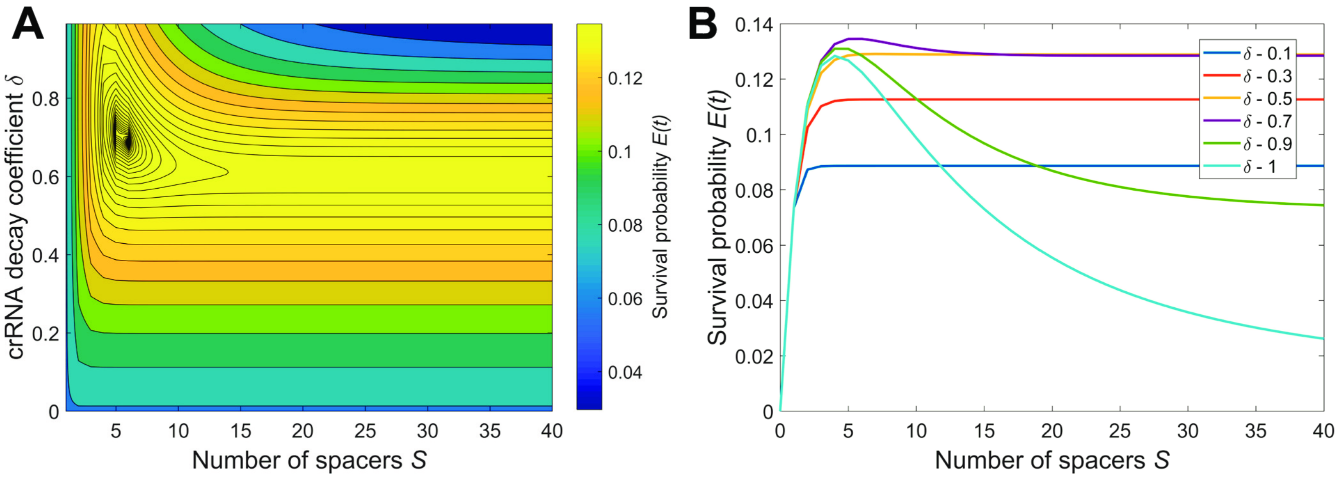

Fig 3. Typical survival probability profile. (A) Plot of survival probability $E(t)$ vs. the crRNA decay coefficient δ and the number of spacers in CRISPR array S. Other parameters are: $\beta = 1$, $\chi = 1.4$, $\mu = 0.9$, and $rNt = 5$. (B) Six curves of $E(t)$ vs. S for various values of δ and same β, χ, m, and rNt as in the panel A.

https://doi.org/10.1371/journal.pcbi.1005891.g003

crRNA with older spacers in effector complexes: no matter how long the array is, only crRNA with the first few spacers are mainly used by effectors. Thus, an "automatic" cutoff in excessive use of older and thus inefficient spacers is implemented.

Naturally, the optimal number of spacers depends on such parameters as protospacer mutation probability $1 - \mu$ and the efficiency of effector binding to its targets β: In Fig 4 we show how the plot of the "typical case" shown above in Fig 3 is affected by changes in these system parameters. An increase in the mutation rate shifts the optimum towards fewer spacers or stronger reliance of the CRISPR-Cas system on crRNA with the first spacer. In the extreme case this can lead to the optimal array containing one spacer only (Fig 4, top-left corner). This corresponds to the case when there is a very high chance that older spacers have mutated, so the benefit from using the second spacer cannot overcome the decrease in the number of effector complexes loaded with crRNA containing the first, most recently acquired spacer. In contrast, an increase of CRISPR interference efficiency shifts the optimum towards more CRISPR spacers and more equal contribution of spacers of different age (Fig 4, bottom-right corner). An increase in the binding efficiency leads to a larger fraction of time the effector spends bound to the protospacer ultimately leading to binding saturation. In this case the sharing of CRISPR effectors between crRNAs with different spacers is beneficial as it allows the effectors to reduce competition for the same protospacer. An increase in the CRISPR interference efficiency χ also leads to an increase in survival probability.

For a more detailed study of the optimal number of spacers, we conducted the following calculations: for each set of "array-independent" parameters μ, β, χ we analyzed the CRISPR efficiency in the whole range of the number of spacers S and crRNA decay coefficients δ. The number of spacers S_{opt} and crRNA decay coefficient δ_{opt} that maximized survival probability, as well as the maximal survival probability itself $E_{max}(t)$ are plotted in Fig 5. As discussed above, higher viral mutation rates lead to lower survival probability and fewer spacers (Fig 5A). For very high mutation probability (above 0.7), the CRISPR interference efficiency approaches zero for all values of other parameters. The mutation rate of viruses caps the CRISPR efficiency as the probability to survive the infection is constrained by the probability

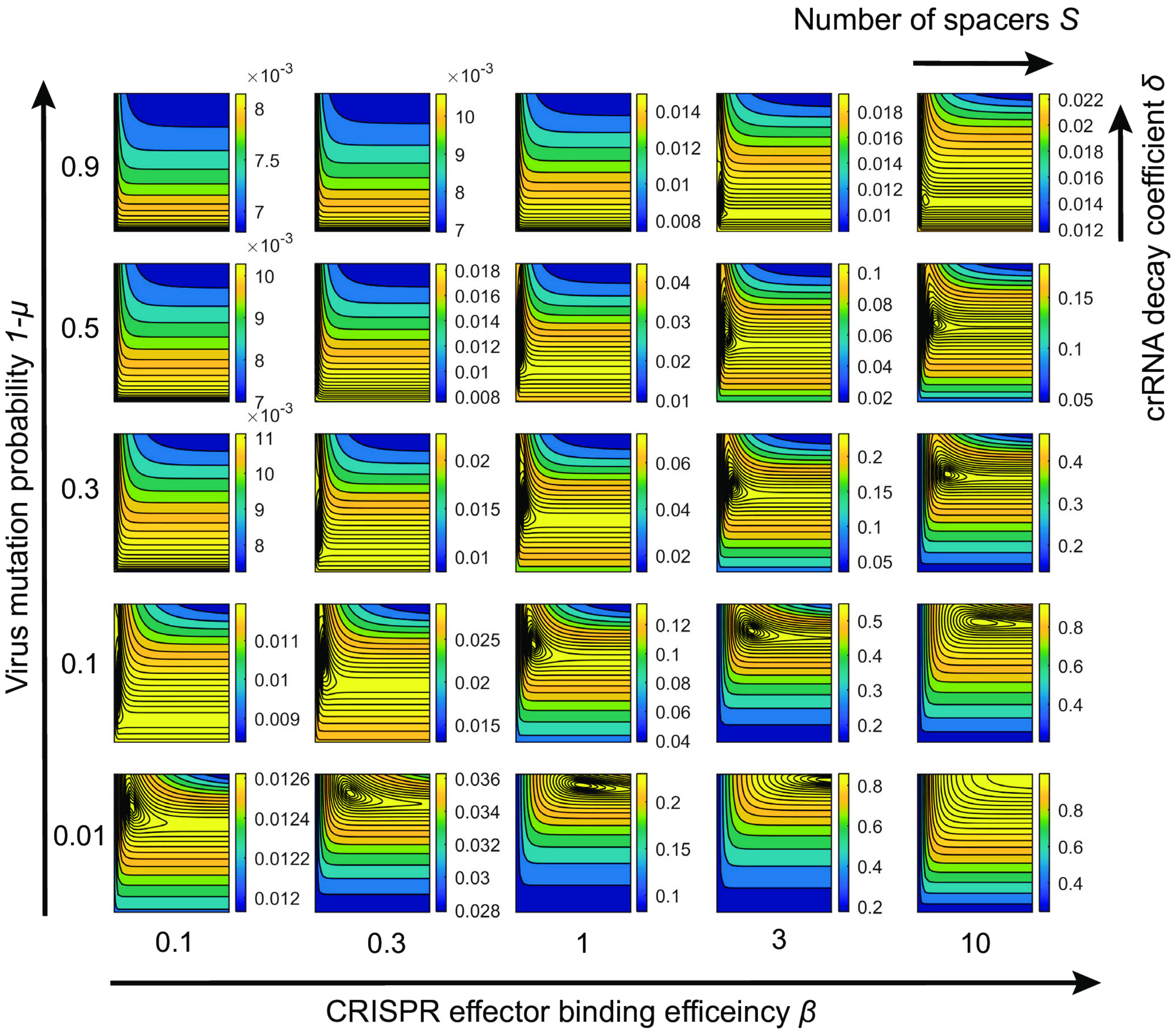

Fig 4. Effects of mutation rate and binding efficiency. A set of 25 panels illustrating how the survival probability depends on S and δ for various values of protospacer mutation probability $1 - \mu$ and binding efficiency of effectors β. The δ and S axes in each small panel have the same range as in the panel A in Fig 3, while the scale of the heat-map varies and is indicated to the right of each panel. The external axes describe the variation of mutation probability $1 - \mu$ and effector binding efficiency β. In all panels $\chi = 1.4$ and $rNt = 5$.

https://doi.org/10.1371/journal.pcbi.1005891.g004

I_{max} that at least one of viral protospacers has not mutated.

$$I_{max} = 1 - \prod_{i=1}^{S} (1 - \mu)^{i-1/2} \tag{17}$$

On the other hand, a high binding β or interference efficiency χ lead to arrays with more spacers and higher survival probability (Fig 5B and 5C). In this case, more CRISPR effectors can complex with crRNAs with older spacers without interfering with the binding to crRNAs with younger spacers due to the system saturation. Arrays with more spacers both increase the

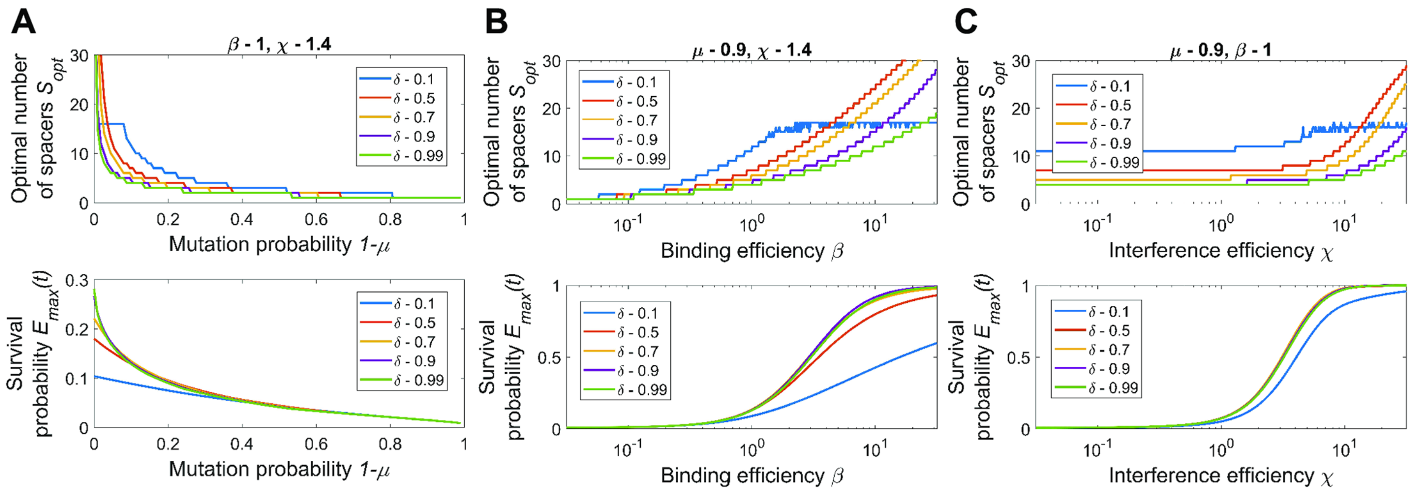

Fig 5. Effect of parameters on the optimal number of spacers and the maximal survival probability. The optimal number of spacers and corresponding survival probability as functions of one of the array-unrelated parameters: (A) As function of mutation probability 1 − *μ*, other parameters are *β* = 1 and *χ* = 1.4. (B) As function of binding efficiency *β*, other parameters are *μ* = 0.9 and *χ* = 1.4. (C) As function of interference efficiency *χ*, other parameters *μ* = 0.9 and *β* = 1. The average number of viral infections was *rNt* = 5 in all panels.

https://doi.org/10.1371/journal.pcbi.1005891.g005

viral DNA degradation rate and, more importantly, reduce the chance that the cell becomes unprotected if some of protospacers mutate. This suggests a correlation between the optimal number of spacers S_{opt} and the maximal protective performance of CRISPR-Cas system $E_{max}(t)$. Comparing the optimal number of spacers and maximal survival probability heatmaps shown in Fig 6, one sees that the parameters that produce high survival probability indeed correspond to arrays with relatively many spacers.

Figs 5 and 6 lead to a conclusion that there is a definite set of parameters for which CRISPR-Cas systems are efficient. The virus mutation probability should remain low on the timescale of spacer acquisition, while the binding of effector complexes to target protospacers and the rate of degradation of viral DNA should be high. This set of parameters favors arrays

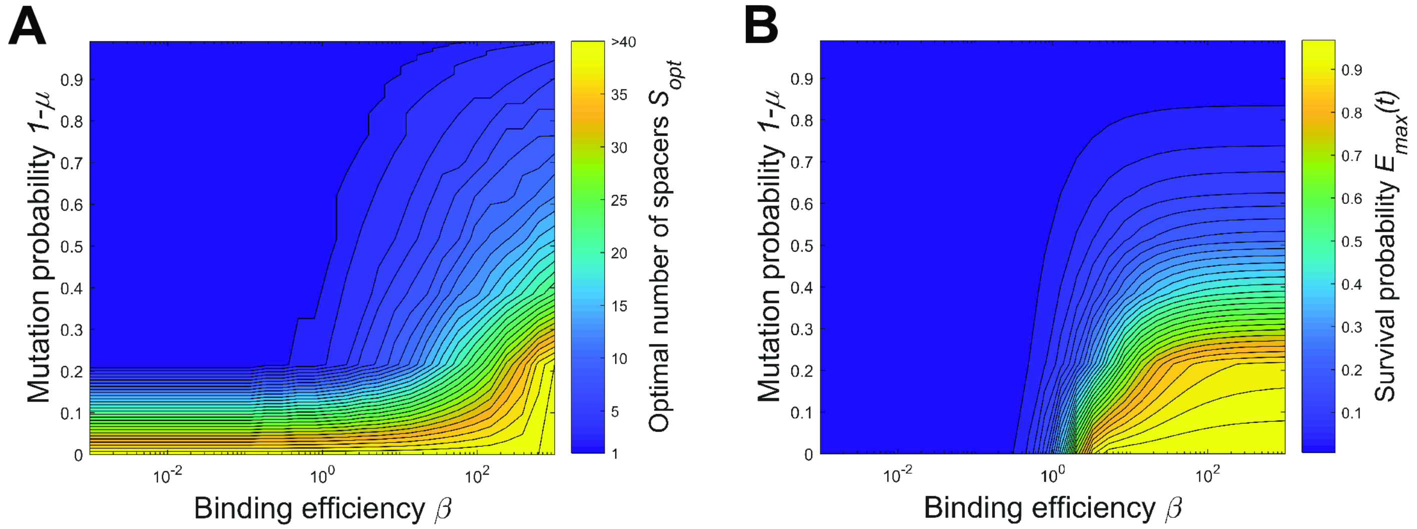

Fig 6. The optimal number of spacers and maximal cell survival probability. The optimal number of spacers (A) and the maximal cell survival probability (B) are shown vs. a range of binding efficiencies *β* and mutation probabilities 1 − *μ* for *rNt* = 5 and *χ* = 1.4.

https://doi.org/10.1371/journal.pcbi.1005891.g006

with more spacers. This can be summarized as a simple rule: Under the conditions that imply high cell survival, the optimal array contains many spacers and is efficient, while under less favorable conditions, the optimal array contains a few (or even one) spacers and is less efficient. In reality, the array composition may change on the timescale of viral infections (for example, via naïve or primed spacer acquisition), which may increase CRISPR interference efficiency by instantaneous insertion of one or a few perfectly matched spacers with high levels of expression of corresponding crRNAs. This, however, goes beyond the important assumption of our model that the array is static on the timescale of viral infection and thus is beyond our present consideration.

Analytical results: Multiple viral species

Consider now a more realistic scenario of a cell confronting several distinct viral species. Using the same logic as in the section above and, specifically considering infections by different viruses being independent of each other, we conclude that the survival probability is given by the Eq (12), where the index of the product j enumerates all viral species, including their mutation variants, present in the system. The interference term associated with a viral species j not targeted by any spacer present in a given array is zero, $I_j = 0$. The corresponding term in the survival probability $\exp(-rNt\nu_j)$ describes the probability for a cell not to encounter such a virus till time t.

Similarly to the case of single viral species, we account for mutation variants of each virus and reduce (12) to the product running over only distinct viral species. In order to simplify further analysis, we denote by ν_i the fraction of the ith virus in the total number of viruses N so that $\nu_i = N_i/N$, where N_i is the number of viral particles of species i. This results in the following expression for survival probability of a cell with a given combination of spacers,

$$E_c(t) = \exp\left[-rNt\sum_{j=1}^{\nu}\nu_j\left(\prod_{i\in\{S_j\}}\{1-m_i[1-\exp(-\chi B_i)]\}\right)\right]. \tag{18}$$

Here the sum over j counts all ν viral species while the product over i enumerates all spacers $\{S_j\}$ taken from the jth virus. As in (15), we approximate m_i by $\mu^{i-1/2}$ assuming again that spacers are acquired in a periodic fashion, with equal times between acquisitions.

The Eq (18) describes survival probability of a cell with a given CRISPR array characterized by sets of spacers $\{S_j\}$ taken from viral species j. In order to evaluate the overall performance of a CRISPR array with S spacers, we need to enumerate survival probabilities for all combination of spacers in such an array. To do so, we assume that the probability to acquire a spacer from a given viral species is proportional to the fraction of such species in the total viral pool. Hence the probability of an array to have a certain combination of spacers is

$$P_c = \prod_{k=1}^{S}\nu_k, \tag{19}$$

where ν_k is the relative concentration of viral species from which the spacer k has been acquired. For example, an array of two spacers (a, b) in a system populated by two viral species 1 and 2 with relative concentrations ν_1 and ν_2 can be in any of the following four forms with corresponding probabilities: $P_{(1,1)} = \nu_1^2$, $P_{(1,2)} = P_{(2,1)} = \nu_1\,\nu_2$, and $P_{(2,2)} = \nu_2^2$.

The average survival probability of a cell in a multiviral medium is a sum of survival probabilities corresponding to each combination of spacers E_c, weighted by the probability to acquire such a combination P_c, and the summation runs over all combinations of spacers.

$$E(t) = \sum_c E_c(t) P_c. \tag{20}$$

Numerical results: Multiple viral species

A typical plot of $E(t)$ is presented in Fig 7. In this calculation we considered two species of viruses with the same population size $v_1 = v_2 = 0.5$. The values of other parameters were the same as in Fig 3: The binding efficiency $\beta = 1$, the interference efficiency $\chi = 1.4$, the probability for a protospacer not to mutate over the typical period between spacer acquisition $\mu = 0.9$, and the typical virus encounter number $rNt = 5$. Comparing to the single-virus case in Fig 3, the total number of viral particles is the same, but the virus pool is now split between two species.

In general, the shape of the survival probability $E(t)$ profile is similar to the single-virus case and $E(t)$ reaches its maximum for a certain crRNA decay coefficient δ and a certain number of spacers S. However, comparing the optimal number of spacers, crRNA decay coefficient, and survival probabilities between the single- and two-virus cases (Figs 3A and 7), one sees that in the two-virus case the maximum is generally shifted towards arrays with more spacers, and $E(t)$ is lower. For a given set of parameters, the addition of the second virus does not significantly shift the optimal S and δ but drops the survival probability dramatically. If the virus mutation rate is lower and the CRISPR interference efficiency is higher, the presence of an

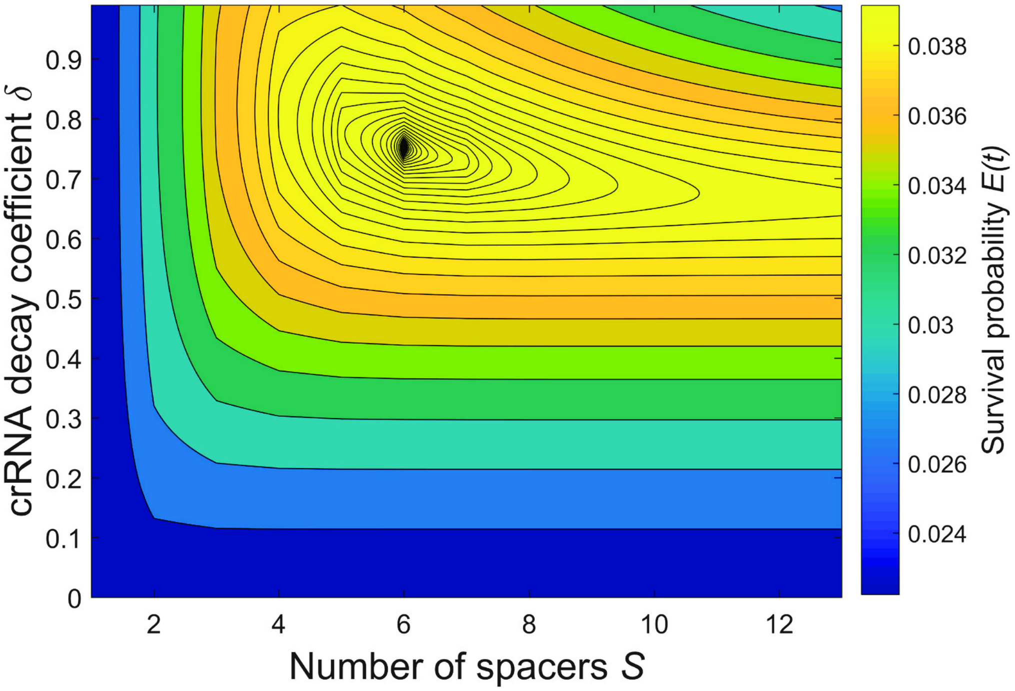

Fig 7. CRISPR performance for two virus species. Plot of the survival probability $E(t)$ as a function of crRNA decay coefficient δ and the number of spacers S of a cell confronting two different viruses with equal population sizes, $v_1 = v_2 = 0.5$. The binding efficiency is $\beta = 1$ and the interference efficiency is $\chi = 1.4$. Viral mutation probability $1 - \mu$ is equal to 0.1 and $rNt = 5$.

https://doi.org/10.1371/journal.pcbi.1005891.g007

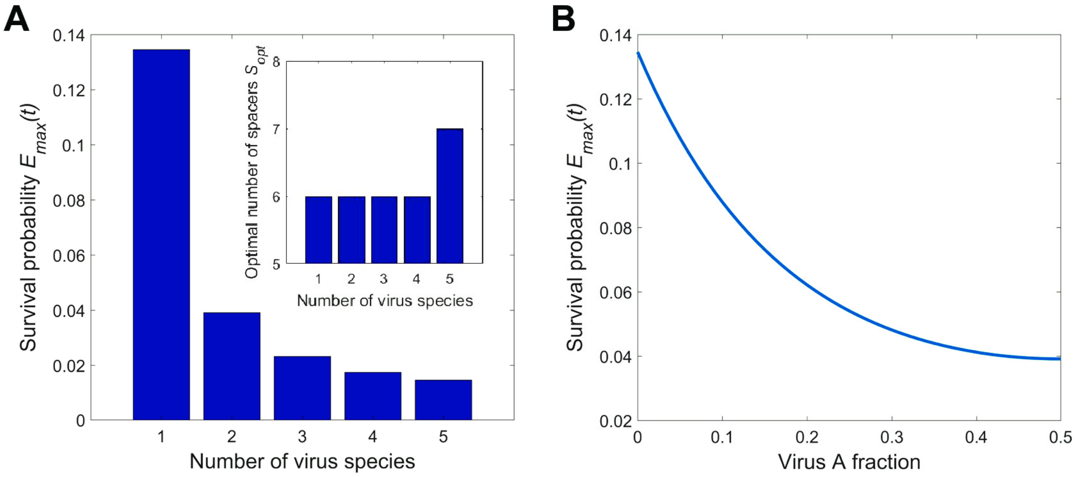

Fig 8. Survival probability vs diversity of the virus pool. Plots of the optimized over δ and S cell survival probability and the number of spacers vs the number of viral species and the composition of a two-virus pool for $\beta = 1$, $\chi = 1.4$, $\mu = 0.9$ and $rNt = 5$. (A) Maximal survival probability $E(t)$ (outer plot) and optimal number of spacers S_{opt} (inner plot) as a function of the number of virus species n. The abundance of virions belonging to different species in the viral pool are the same for all species, $v_1 = \ldots = v_n = 1/n$. (B) The maximal survival probability vs the relative abundance of one of the viruses in a two-virus pool.

https://doi.org/10.1371/journal.pcbi.1005891.g008

additional viral species will affect the optimal S and δ more strongly. However, relating the model parameters to the experimental results [35], it is unlikely that the CRISPR efficiency can be significantly higher in vivo than the numbers shown in Fig 7.

When the number of virus species in the total virus pool increases even without a change in the total viral particles concentration, the survival probability approaches zero (Fig 8A). This occurs because the efficient number of spacers is limited by the virus mutation rate and the number of effector complexes present in the cell (encoded in the coefficient β). In other words, further increase in the number of spacers does not lead to any increase in protective function of CRISPR-Cas. Since an array of an effectively limited number of spacers has to contain spacers from more virus species, fewer spacers match each virus and the survival probability decreases.

Another observation is obtained considering the two-virus case and changing the ratio of those viruses in the pool (Fig 8B). As expected, the survival probability reaches a maximum when the fraction of one virus approaches zero (which correspond to the single-virus case) and goes to a minimum when both viruses are equally abundant.

This brings us to the conclusion that survival probability of a cell dramatically depends on the diversity of the viral pool.

Discussion

The function of CRISPR-Cas as prokaryotic adaptive immune system has been extensively studied from the point of view of molecular mechanisms. Its ecological role and its contribution to the "arms race" between prokaryotes and their viruses have been analyzed in many evolutionary dynamics models and found to be very complex and often unpredictable. In this work, we qualitatively explored the forces affecting the number of spacers in a CRISPR array. We found that more spacers in a CRISPR array targeting a virus decrease the chances of the virus to escape detection through simultaneous mutation in all targeted protospacers. Also, more spacers lead to more effective use of CRISPR effectors, distributing them between a larger number of target protospacers, which results in higher probability of viral DNA destruction. However, at the same time, more diverse crRNA repertoire results in fewer effector complexes charged with crRNAs containing recently acquired spacers that target protospacers least likely to mutate. This "dilution" effect agrees with recent experimental results, showing that removal of non-matching spacers from the array can lead to a dramatic increase in interference efficiency by remaining spacers [38].

The interplay of described forces leads to the optimum in the number of spacers per array, determined by the properties of the CRISPR-Cas system and the diversity and mutation rates of viral species in the following way: A better binding of the CRISPR effectors to their targets and faster rate of target DNA degradation allow a prokaryotic cell to maintain more spacers in the array and increase its survival probability. Also, less frequent mutations in viral protospacers create an opportunity for hedging against those mutations by keeping more of previously acquired spacers. In contrast, a less efficient kinetics of binding and viral DNA cutting and faster-mutating viruses make arrays with fewer spacers more advantageous.

We consider this work to be a necessarily conceptual study of optimality of CRISPR arrays. However, while the final results of our analysis presented in subsections "Numerical results: Single viral species" and "Numerical results: Multiple viral species" are applicable only to a particular ("average") set of virus-host coexistence scenarios, our more general estimates for the survival probability given in Eqs (12 and 15) can be used as building blocks in more complex and hopefully more accurate dynamical models. A few additional comments on the applicability of our results and biological insights that can follow from them are in order.

Deviations from steady state

Our results were derived explicitly assuming a steady state of the CRISPR-virus dynamics. However, in previous research, both modeling and experimental, it was shown that CRISPR systems are far from being stable, undergoing periodic and irregular variations that play an important role in their function [21, 39]. While in our analysis we assumed that the viral environment (i.e. species composition and concentrations) is constant (except for appearance of mutant protospacers), the actual viral dynamics, which is commonly non-steady, may affect the optimal number of spacers in CRISPR arrays. It is important to note that the number of spacers providing the maximum defensive efficiency of CRISPR-Cas system and maximum cell survivability is mechanistically achieved through the evolution of rates of acquisition and loss of spacers. Any combination of spacer acquisition and loss rates would result in a steady state, which, in the first approximation, is controlled by the ratio of the former and the latter. The time to reach this steady state can be estimated roughly as the inverse of the spacer acquisition rate times the steady state number of spacers. However, these factors change both due to variations in the ecological environment (frequency and mutation diversity of viral infections), and because of the evolution of the CRISPR machinery itself. Thus we see this process in dynamics: spacer uptake and loss rates determine steady state number of spacers and rates are being evolved in order to reach optimal steady state number of spacers for the given environment.

For an incredible diversity of possible forms of viral-host coexistence scenarios, the time scale of changes in the viral environment varies enormously and presumably can be very low, allowing the optimal number of spacers to accumulate in an almost steady ecological environment. In the opposite limit of much slower than population dynamics spacer acquisition, the array content represents some average and perhaps delayed sample of the viral pool and the function of CRISPR system is generally suboptimal. It is also appealing to speculate that the observed coexistence of several types of CRISPR systems in the same prokaryotic genome has evolved as a way to optimize the immune response to several quite distinct types of viral environment with different dynamic timescales.

At the same time, one could imagine ecological conditions when the spacer uptake and loss independently (rather than via their ratio) affect the number of spacers in the array. For instance, an increase in both the acquisition and loss rates, which keeps their ratio constant, would nevertheless lead to a gradual depletion of spacers if viral attacks are so infrequent that new spacers are nowhere to come from. In such scenarios, the observed number of spacers can be drastically different from our predictions.

Since the expression of *cas* genes is likely to be regulated and in some cases can be turned up by viral invasions [32, 33], the question arises of how non-stationarity in the level of Cas proteins affects our conclusions about the optimal number of spacers. Using the same approach, one could generalize our results to account for time dependence of Cas protein levels over a course of viral attack. This will lead to a more complex expression for the interference probability, which will depend on generally not quantitatively-understood kinetics of both virus attack and CRISPR-Cas system activation. Our results were computed using the data for the constitutive expression of *cas* genes [35], thus presenting an upper bound for the survival probability. In principle, one can use the time average of the number of Cas protein complexes as C in the expression for the binding efficiency β (5) to get the best approximate estimate for the efficiency of CRISPR defence and the optimal number of spacers.

Our assumption that all protospacers have equal probability to mutate is definitely not universal. It has been observed [13, 40] and modeled [41] that older spacers often correspond to evolutionary-conserved regions in viral genomes, causing higher survival rates during infection by preventing formation of viral escape mutants, thus, explaining the ubiquitous presence of their bearers. In the framework of our model, this can be taken into account by assigning individual values to the probabilities for a protospacer i to stay mutation-free m_i in Eq (15). The resulting expressions for the interference probability can be used in more complex evolutionary and population dynamics models to study the evolution of the spacer content.

Comparison with existing models

Our results generally agree with the main findings of models existing in the field: We confirm that a higher diversity of viral environment results in a dominance of viruses over the CRISPR system [22, 42]. This effect could be achieved by either a high number of virus species in the environment or a high mutation rate of viruses belonging to the single species (often associated with large viral population). However, here we have also shown that a diversity of virus species leads to arrays with more spacers while a higher viral mutation rate leads to arrays with fewer spacers. This agrees with a proposed hypothesis that a lower viral mutation rate leads to arrays with on average more spacers in thermophilic bacteria [42]. Another important note on comparing our model with existing ones is related to the definition of probability of CRISPR immunity failure. Some of the models used a binary approach to immunity failure [21]. Either the infected cell kills the virus or the virus kills the cell and reproduces normally. We define the CRISPR failure probability $1 - I$ as the probability of viral DNA not getting cut by CRISPR effectors/executors during viral DNA duplication cycle. Distinguishing between these two approaches is important as it affects the interpretation of parameters obtained from experiments. For example, a CRISPR-Cas system can remain active in doomed or dead cells, resulting in lower viral burst size and fewer secondary infections [35]. Our analysis based on [35] (S2 Appendix) resulted in the estimate of the CRISPR failure probability around 30% compared to 10^{-5} in [21].

Importance of palindromic nature of CRISPR repeats

One of important observations is that the equipartition of crRNA between CRISPR effector complexes is not optimal and a decrease of the fraction of older crRNA bound to effectors increases the overall efficiency of the immune response. While there is a limited pool of effectors, they serve better when binding to crRNAs with most recently acquired spacers. Since the probability that a spacer no longer matches the protospacer increases with time, Cas effectors should either have a higher affinity towards crRNA from younger spacers (which is impossible to accomplish) or crRNA containing more recent spacers should be more abundant. This latter may be implemented naturally owing to formation of hairpin by CRISPR repeats in the primary array transcripts [43, 44]. It is well known that hairpins have a potential to pause or terminate transcription elongation [45, 46]. The longer the array is, the more hairpins need to be transcribed and the higher the chance is that transcription would be terminated before the RNA polymerase reaches the end of the array. This could result in more abundant shorter pre-crRNAs that include only the younger spacers. At the same time, certain CRISPR repeats are found to be only weakly palindromic, such as those in type II CRISPR systems [47].

Another possible mechanism to control the abundance of crRNA derived from newer and older spacers is binding of regulatory proteins that specifically target CRISPR repeats [48]. If these proteins act as transcription terminators, such binding also results in exponential-like distribution of spacers.

Fitness cost of CRISPR system

While in our study we ignored the fitness costs of an active CRISPR system, we find it important to discuss it as these were studied in various experimental works and included in some models [49]. It has been shown in a number of publications that the activity of CRISPR systems is under strong evolutional pressure. There are various factors that can contribute to the cost of CRISPR including genomic burden [50], the cost of maintenance of cas genes [19], self-immunity [51] and blockage of beneficial horizontal gene transfer (HGT) [17]. However genomic burden seems not to be significant in most cases as even the largest of the CRISPR systems contribute only 1% to the total size of a prokaryotic genome [11]. In the case of self-immunity, it seems to be related to the very process of acquisition of new spacers, thus, self-immunity only indirectly affects the number of spacers in CRISPR array [52–54]. For the cost of gene maintenance [19] and blockage of HGT [20], it has been shown that an increase in the number of spacers also does not have any significant fitness cost. Thus, in this work, we considered that the fitness cost of CRISPR system did not affect the optimal number of spacers in CRISPR array. In other words, there is no additional fixed cost of the spacer apart from the one arising from Cas effector dilution. That resulted in separation of the number of spacers question from the overall fitness. The factors described in this work affect the optimum number of spacers in CRISPR array and the total fitness benefit of CRISPR system. And this total fitness benefit now can be compared to the fitness cost of CRISPR-Cas system maintenance, that will give the answer whether the CRISPR system will be effective or tends to be knocked out [55].

Primed adaptation in the framework of the model

In this work we have only considered arrays produced in course of naïve, or completely random and relatively infrequent adaptation. Yet it is possible to qualitatively access the effect of primed adaptation on cell survival. Primed adaptation is extremely efficient compared to naïve adaptation since the uptake of spacers happens on the timescale of viral attacks [34]. Its effect on cell survival is at least two-fold. First, there is a direct increase in cell survival probability, which happens when otherwise doomed cells with a non-perfect match between spacers and corresponding protospacers survive the attack by quickly acquiring new spacers. In the first approximation, this effect can be taken into account by rescaling (increasing) the probabilities μ for protospacers to remain mutation-free. Second, the spacer acquisition is no longer controlled only by the viral environment, but also by the presence of particular spacers, which prime adaptation, in the array. This makes the array content highly correlated and makes it impossible to apply our model for multiple viruses. However, in the single-virus case, when all spacers come from the same virus anyway, the primed adaptation simply means that the virus mutation probability $1 - \mu$ becomes very low. Another peculiar feature of primed adaptation is that more than one spacer can simultaneously be taken from the same virus. This results in the series of spacers that get the same probability of a mismatch in further course of the evolution.

Evidently, the primed adaptation improves cell survival during infection. However, apart from an apparent increase in the optimal number of spacers due to a larger effective μ (Fig 5A), it appears impossible without a thorough quantitative study to make a more detailed prediction of how primed adaptation would affect the optimal number of spacers.

Abortive infections and altruistic behavior

In addition to providing immunity and thus saving an infected cell, CRISPR system also "altruistically" decreases the number of secondary infections, originating from infected cell [36, 56], reducing the virus burst size (number of progeny viruses) [35, 57]. This constitutes the second source of selection pressure on the CRISPR functioning.

We analyzed how to minimize the viral burst in section S1 Appendix. It appears that the condition for the minimum of the viral burst (S5) is similar to that for cell survival, (15), but with the rescaled interference efficiency, $\chi' = \nu\chi$. Here $\nu \approx 6 - 7$ is the average number of virus replications in a CRISPR-free cell. This condition leads to the optimal number of spacers which is a bit larger than that for cell survival (Fig 5C and S1 Appendix).

In reality, the optimal number of spacers is somewhere in between those determined for χ and for $\chi' \approx 7\chi$. It is impossible to give a more precise answer as these two optima are often under different types of selection pressures: In the environment with low host cell density, survival of each cell is important while the probability of secondary infection is small. In contrast, when the host cell density is high, it is evolutionary more beneficial to sacrifice a few individual cells but to limit the number of secondary infections.

Conclusions

- We theoretically predict the optimal number of spacers in a CRISPR array which falls into reasonable range from the viewpoint of current experimental data and show that it depends on the interference efficiency of CRISPR effector, crRNA spacer-protospacer binding efficiency, and virus mutation rate.
- Good (from the "point of view" of the cell) conditions, such as high interference and binding efficiencies and slow mutation of viral protospacers, favor arrays with more spacers, which provide better immune protection. Conversely, less favorable conditions shift the optimum to arrays with fewer spacers and less efficient immune protection.
- The majority of optimal array configurations have a non-uniform distribution of unique crRNAs among CRISPR effector complexes with a preference for crRNAs with more recently acquired spacers.

- Fighting against multiple viral species shifts the optimum towards arrays with more spacers and dramatically decreases the maximum efficiency of the CRISPR system.

Supporting information

S1 Appendix. CRISPR-induced reduction in the viral burst.
(PDF)

S2 Appendix. Calculation of the CRISPR interference efficiency from experimental data.
(PDF)

Acknowledgments

We thank Ekaterina Semenova for bringing to our attention several references and sharing the experience from her extensive experimental studies of CRISPR systems.

Author Contributions

Conceptualization: Iaroslav Ispolatov.

Formal analysis: Alexander Martynov, Iaroslav Ispolatov.

Investigation: Alexander Martynov, Konstantin Severinov, Iaroslav Ispolatov.

Methodology: Alexander Martynov, Konstantin Severinov, Iaroslav Ispolatov.

Software: Alexander Martynov.

Supervision: Konstantin Severinov, Iaroslav Ispolatov.

Visualization: Alexander Martynov.

Writing – original draft: Alexander Martynov, Konstantin Severinov, Iaroslav Ispolatov.

Writing – review & editing: Alexander Martynov, Konstantin Severinov, Iaroslav Ispolatov.

References

1. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology direct. 2006; 1(1):7. https://doi.org/10.1186/1745-6150-1-7 PMID: 16545108
2. Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005; 151(8):2551–2561. https://doi.org/10.1099/mic.0.28048-0 PMID: 16079334
3. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007; 315(5819):1709–1712. https://doi.org/10.1126/science.1138140 PMID: 17379808
4. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nature reviews Microbiology. 2011; 9(6):467–77. https://doi.org/10.1038/nrmicro2577 PMID: 21552286
5. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology. 2015; 13(11):722–736. https://doi.org/10.1038/nrmicro3569 PMID: 26411297
6. Hargreaves KR, Flores CO, Lawley TD, Clokie MRJ. Abundant and Diverse Clustered Regularly Interspaced Short Palindromic Repeat Spacers in Clostridium difficile Strains and Prophages Target Multiple Phage Types within This Pathogen. mBio. 2014; 5(5):e01045–13–e01045–13. https://doi.org/10.1128/mBio.01045-13 PMID: 25161187

7. McGhee GC, Sundin GW. Erwinia amylovora CRISPR elements provide new tools for evaluating strain diversity and for microbial source tracking. PLoS ONE. 2012; 7(7). https://doi.org/10.1371/journal.pone.0041706

8. van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras Jb, Barbu EM, et al. Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant Pseudomonas aeruginosa. mBio. 2015; 6(6):1–13. https://doi.org/10.1128/mBio.01796-15

9. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Research. 2013; 41(8):4360–4377. https://doi.org/10.1093/nar/gkt157 PMID: 23470997

10. Agari Y, Sakamoto K, Tamakoshi M, Oshima T, Kuramitsu S, Shinkai A. Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection. Journal of Molecular Biology. 2010; 395 (2):270–281. https://doi.org/10.1016/j.jmb.2009.10.057 PMID: 19891975

11. Rath D, Amlinger L, Rath A, Lundgren M. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie. 2015; 117:119–128. https://doi.org/10.1016/j.biochi.2015.03.025 PMID: 25868999

12. Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJM. Diversity of CRISPR loci in Escherichia coli. Microbiology. 2010; 156(5):1351–1361. https://doi.org/10.1099/mic.0.036046-0

13. Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. Journal of Bacteriology. 2008; 190 (4):1401–1412. https://doi.org/10.1128/JB.01415-07 PMID: 18065539

14. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC bioinformatics. 2007; 8:172. https://doi.org/10.1186/1471-2105-8-172 PMID: 17521438

15. Hale C, Kleppe K, Terns RM, Terns MP. Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. RNA (New York, NY). 2008; 14(12):2572–9. https://doi.org/10.1261/rna.1246808

16. Levin BR, Moineau S, Bushman M, Barrangou R. The Population and Evolutionary Dynamics of Phage and Bacteria with CRISPR-Mediated Immunity. PLoS Genetics. 2013; 9(3). https://doi.org/10.1371/journal.pgen.1003312

17. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008; 322(5909):1843–1845. https://doi.org/10.1126/science.1165771 PMID: 19095942

18. Bondy-Denomy J, Davidson AR. To acquire or resist: The complex biological effects of CRISPR-Cas systems. Trends in Microbiology. 2014; 22(4):218–225. https://doi.org/10.1016/j.tim.2014.01.007 PMID: 24582529

19. Vale PF, Lafforgue G, Gatchitch F, Gardan R, Moineau S, Gandon S. Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. Proceedings of the Royal Society B: Biological Sciences. 2015; 282(1812):20151270. https://doi.org/10.1098/rspb.2015.1270 PMID: 26224708

20. Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. The ISME journal. 2015; 9(9):2021–7. https://doi.org/10.1038/ismej.2015.20 PMID: 25710183

21. Childs LM, Held NL, Young MJ, Whitaker RJ, Weitz JS. Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. Evolution. 2012; 66(7):2015–2029. https://doi.org/10.1111/j.1558-5646.2012.01595.x PMID: 22759281

22. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. Journal of Bacteriology. 2013; 195(17):3834–3844. https://doi.org/10.1128/JB.00412-13 PMID: 23794616

23. Bradde S, Vucelja M, Tesileanu T, Balasubramanian V. Dynamics of adaptive immunity against phage in bacterial populations. PLOS Computational Biology. 2017; 13(4):e1005486. https://doi.org/10.1371/journal.pcbi.1005486 PMID: 28414716

24. Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJM. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biology. 2013; 10(5):792–802. https://doi.org/10.4161/rna.24023 PMID: 23445770

25. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJJ. CRISPR-Cas: Adapting to change. Science. 2017; 356(6333):eaal5056. https://doi.org/10.1126/science.aal5056 PMID: 28385959

26. Zoephel J, Randau L. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. Biochemical Society Transactions. 2013; 41(6):1459–1463. https://doi.org/10.1042/BST20130129 PMID: 24256237

27. Semenova E, Jore MM, Datsenko Ka, Semenova A, Westra ER, Wanner B, et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(25):10098–10103. https://doi.org/10.1073/pnas.1104144108 PMID: 21646539

28. Fischer S, Maier LK, Stoll B, Brendel J, Fischer E, Pfeiffer F, et al. An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. Journal of Biological Chemistry. 2012; 287(40):33351–33365. https://doi.org/10.1074/jbc.M112.377002 PMID: 22767603

29. Shah S, Erdmann S, Mojica F, Garrett R. Protospacer recognition motifs. RNA biology. 2013; 10(May):891–899. https://doi.org/10.4161/rna.23764 PMID: 23403393

30. Pul Ü, Wurm R, Arslan Z, Geißen R, Hofmann N, Wagner R. Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. Molecular Microbiology. 2010; 75(6):1495–1512. https://doi.org/10.1111/j.1365-2958.2010.07073.x PMID: 20132443

31. Westra ER, Pul Ü, Heidrich N, Jore MM, Lundgren M, Stratmann T, et al. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Molecular Microbiology. 2010; 77(6):1380–1393. https://doi.org/10.1111/j.1365-2958.2010.07315.x PMID: 20659289

32. Liu T, Li Y, Wang X, Ye Q, Li H, Liang Y, et al. Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Research. 2015; 43(2):1044–1055. https://doi.org/10.1093/nar/gku1383 PMID: 25567986

33. Stratmann T, Pul Ü, Wurm R, Wagner R, Schnetz K. RcsB-BglJ activates the Escherichia coli leuO gene, encoding an H-NS antagonist and pleiotropic regulator of virulence determinants. Molecular Microbiology. 2012; 83(6):1109–1123. https://doi.org/10.1111/j.1365-2958.2012.07993.x PMID: 22295907

34. Semenova E, Savitskaya E, Musharova O, Strotskaya A, Vorontsova D, Datsenko KA, et al. Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113(27):7626–7631. https://doi.org/10.1073/pnas.1602639113 PMID: 27325762

35. Strotskaya A, Savitskaya E, Metlitskaya A, Morozova N, Datsenko KA, Semenova E, et al. The action of Escherichia coli CRISPR–Cas system on lytic bacteriophages with different lifestyles and development strategies. Nucleic Acids Research. 2017;(15):gkx042. https://doi.org/10.1093/nar/gkx042

36. Westra ER, Buckling A, Fineran PC. CRISPR-Cas systems: beyond adaptive immunity. Nature reviews Microbiology. 2014; 12(5):317–26. https://doi.org/10.1038/nrmicro3241 PMID: 24704746

37. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal. 2011; 30(7):1335–1342. https://doi.org/10.1038/emboj.2011.41 PMID: 21343909

38. Rao C, Chin D, Ensminger AW. Priming In A Permissive Type I-C CRISPR-Cas System Reveals Distinct Dynamics Of Spacer Acquisition And Loss. RNA. 2017; p. rna.062083.117.

39. Berezovskaya FS, Wolf YI, Koonin EV, Karev GP. Pseudo-chaotic oscillations in CRISPR-virus coevolution predicted by bifurcation analysis. Biology direct. 2014; 9(1):13. https://doi.org/10.1186/1745-6150-9-13 PMID: 24986220

40. Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, et al. The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Molecular Microbiology. 2012; 85(6):1057–1071. https://doi.org/10.1111/j.1365-2958.2012.08172.x PMID: 22834929

41. He J, Deem MW. Heterogeneous Diversity of Spacers within CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). Physical Review Letters. 2010; 105(12):128102. https://doi.org/10.1103/PhysRevLett.105.128102 PMID: 20867676

42. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV. Viral diversity threshold for adaptive immunity in prokaryotes. mBio. 2012; 3(6):1–10. https://doi.org/10.1128/mBio.00456-12

43. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome biology. 2007; 8(4):R61. https://doi.org/10.1186/gb-2007-8-4-r61 PMID: 17442114

44. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (New York, NY). 2008; 321(5891):960–4. https://doi.org/10.1126/science.1159689

45. Wilson KS, Hippel PHV. Transcription termination at intrinsic terminators: The role of the RNA hairpin (Escherichia coli/RNA polymerase/rho-independent termination). Biochemistry. 1995; 92(September):8793–8797.

46. Farnham PJ, Platt T. Rho-independent termination: Dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. Nucleic Acids Research. 1981; 9(3):563–577. https://doi.org/10.1093/nar/9.3.563 PMID: 7012794

47. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Research. 2014; 42(10):6091–6105. https://doi.org/10.1093/nar/gku241 PMID: 24728998

48. Deng L, Kenchappa CS, Peng X, She Q, Garrett RA. Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. Nucleic Acids Research. 2012; 40(6):2470–2480. https://doi.org/10.1093/nar/gkr1111 PMID: 22139923

49. Han P, Deem MW. Non-classical phase diagram for virus bacterial coevolution mediated by clustered regularly interspaced short palindromic repeats. Journal of The Royal Society Interface. 2017; 14 (127):20160905. https://doi.org/10.1098/rsif.2016.0905

50. Kuo CH, Ochman H. Deletional Bias across the Three Domains of Life. Genome Biology and Evolution. 2010; 1:145–152. https://doi.org/10.1093/gbe/evp016

51. Vercoe RB, Chang JT, Dy RL, Taylor C, Gristwood T, Clulow JS, et al. Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. PLoS Genetics. 2013; 9(4). https://doi.org/10.1371/journal.pgen.1003454 PMID: 23637624

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

53. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015; 520(7548):505–510. https://doi.org/10.1038/nature14302 PMID: 25874675

54. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research. 2012; 40(12):5569–5576. https://doi.org/10.1093/nar/gks216 PMID: 22402487

55. Jiang W, Maniv I, Arain F, Wang Y, Levin BR, Marraffini LA. Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. PLoS Genetics. 2013; 9(9):e1003844. https://doi.org/10.1371/journal.pgen.1003844 PMID: 24086164

56. Makarova KS, Anantharaman V, Aravind L, Koonin EV. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biology Direct. 2012; 7(1):40. https://doi.org/10.1186/1745-6150-7-40 PMID: 23151069

57. Severinov K, Ispolatov I, Semenova E. The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. Frontiers in Molecular Biosciences. 2016; 3. https://doi.org/10.3389/fmolb.2016.00045 PMID: 27630990