Optimal number of spacers in CRISPR arrays

Alexander Martynov[1*], Konstantin Severinov[1,2,3], Yaroslav 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, NJ, USA

3 Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia

4 Department of Physics, University of Santiago de Chile, Santiago, Chile

* YI jaros007@gmail.com, AM alexander.martynov@skolkovotech.ru

Abstract

We estimate the number of spacers in a CRISPR array of a bacterium which maximizes its protection against a viral attack. The optimality follows from a competition between two trends: too few distinct spacers make the bacteria 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 effective 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 array length lies within a range quite reasonable from the viewpoint of recent experiments.

Author summary

CRISPR-Cas system is an adaptive immunity defense in bacteria and archaea against viruses. It works by accumulating in bacterial genome an array of spacers, or fragments of virus DNA from

previous attacks. By matching spacers to corresponding parts of virus DNA called protospacers, CRISPR-Cas system identifies and destroys intruder DNA. Here we theoretically estimate the number of spacers that maximizes bacterial survival. This optimum emerges from a competition between two trends: More spacers allow a bacterium to hedge against mutations in viral protospacers. However, keeping too many spacers makes the older ones inefficient because of accumulation of mutations in corresponding protospacers in viruses. Thus, fewer CRISPR-Cas molecular machines are left armed with more efficient young spacers. We have shown that a higher efficiency of CRISPR-Cas system allows a bacterium to utilize more spacers, increasing the optimal array length. 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. We think that our study provides a new viewpoint at a huge variety in the observed array lengths and adds relevance to evolutionary models of bacterial-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, numbers and origin 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 bacterial 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 bacterial 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 bacterium (rather than, for example, the survival of a bacterial species) from viruses? We show that the number of CRISPR spacers is primarily limited by dilution" of CRISPR complexes carrying the most immune-active 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. Another simplifying assumption we made is that instead of focusing on the actual evolution that occurs in ever-changing natural viral and bacterial 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 bacterial survival chances.

THE MODEL

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 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]. 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. 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 most recently acquired and thus most efficient spacers, replacing them with older spacers whose com-

plimentary protospacers had a longer time to accumulate mutations. For simplicity, we assume that a single mismatch between a spacer and its protospacer makes the spacer ineffective [3]. While the reality is more complex and certain mutations in a protospacer do not preclude its recognition by the spacer [26], mutations in protospacer adjacent motif [27, 28] or seed region [26]) 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 [26–28].
- 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 effector complexes in the cell is constant. While there is some evidence that cas genes expression might be regulated in vivo [29], the assumption that at given conditions expression levels are constant seems to be reasonable.
- There is only a single copy of viral DNA inside the cell upon infection, i.e., the multiplicity of infections is low.
- CRISPR arrays have a constant size and composition that do not change on the timescale of viral infection, i.e., there is no CRISPR adaptation.
- A single encounter between CRISPR-effector and virus resolves on a shorter timescale than the time between subsequent encounters.
- We do not take into account any fitness costs of maintaining an array of a given length [19, 20].

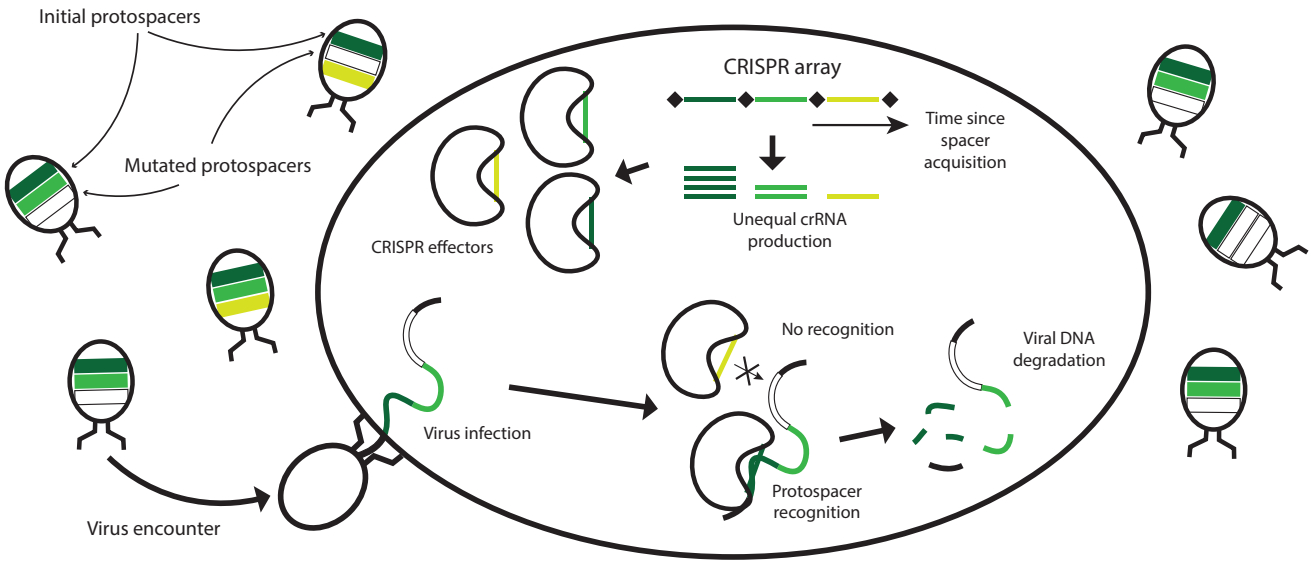

FIG. 1: **Functioning of CRISPR system.** Three spacers are colored according to their age from the time of their acquisition, from dark green marking the youngest spacers to yellow marking the oldest one. Phages carry protospacers colored similarly to their matching spacers; mutated protospacers are colored white. There are more mutated protospacers among older protospacers than among the younger ones. Inside the cell, bean-shaped objects are CRISPR effector complexes armed with individual crRNAs. Complexes with 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.

Probability of interference

Assume that a cell carries an array consisting of S 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

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

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 the age 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 [30]. 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

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

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

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

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

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

Constants β and χ have simple interpretations in terms of familiar Michaelis kinetics: The process of interference can be viewed as a transformation of viral DNA catalyzed by CRISPR effectors. The binding efficiency β corresponds to the inverse Michaelis constant per substrate concentration, S/K_M in standard notations, (6,9). The interference efficiency χ corresponds to the maximal reaction rate in Michaelis kinetics, i.e. the rate at which viral DNA bound by CRISPR effector is degraded.

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.

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 $m_1 m_2 N$, 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 $m_2(1-m_1)N$, 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

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

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,

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

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

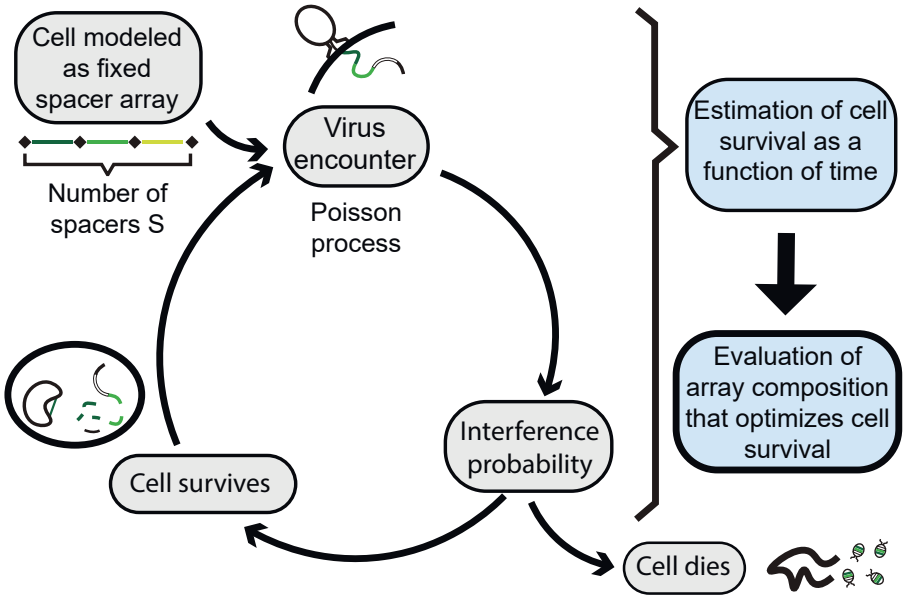

FIG. 2: **Scheme of calculations.** A cell with $S = 3$ CRISPR spacers encounters viruses as a Poisson process with an average rate rN. During each encounter there is either a successful interference with probability I or the cell dies with probability $1 - I$. We evaluate the probability $E(t)$ of the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

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),

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

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.

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. We inferred the interference probability $I_1 \approx 0.7$ of a

single spacer array from the experimental data [31] and set the binding efficiency $\beta = 1$ and the interference efficiency $\chi = 2$ to reproduce the measured single-spacer interference probability. 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 $\delta \approx 0.7$ and $S = 6$. In panel B the dependence of $E(t)$ vs. S is shown for several values of d. Curiously, for low d, the survival $E(t)$ does not noticeably decrease for large S. It happens because of the exponential suppression in frequencies of 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 is implemented.

Naturally, the optimal number of spacers depends on such parameters as protospacer mutation rate $1 - \mu$ 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 (data not shown).

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 mu-

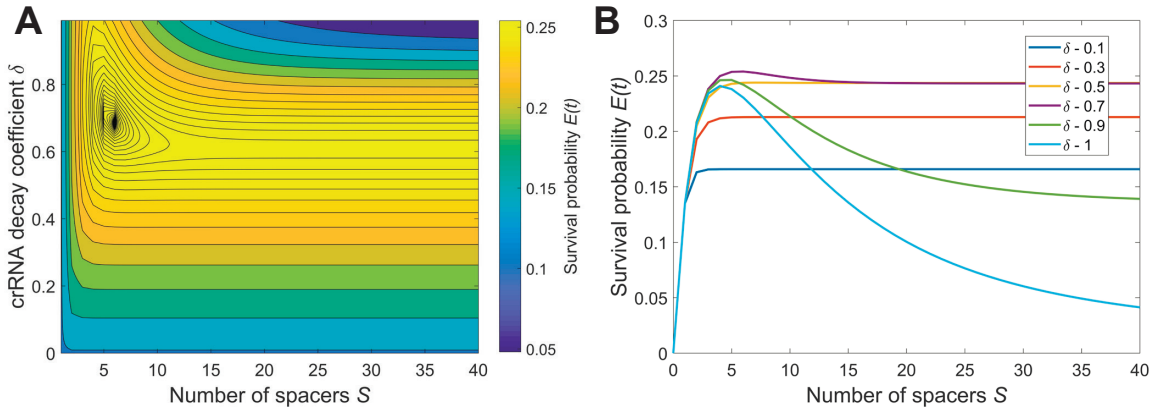

FIG. 3: **Typical survival probability profile.** (A) Plot of survival probability $E(t)$ vs. the crRNA decay coefficient δ and the number of spacers in CRISPR array S. Other parameters are: $\beta = 1$, $\chi = 2$, $\mu = 0.9$, and $rNt = 5$. (B) Six curves of $E(t)$ vs. S for various values of δ and same β, χ, m, and rNt as in the panel A.

tation 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 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, C). 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 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 heat-maps shown in Fig. 6, one sees that the parameters that produce high survival probability indeed correspond to arrays with relatively many (more than 10 spacers) 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

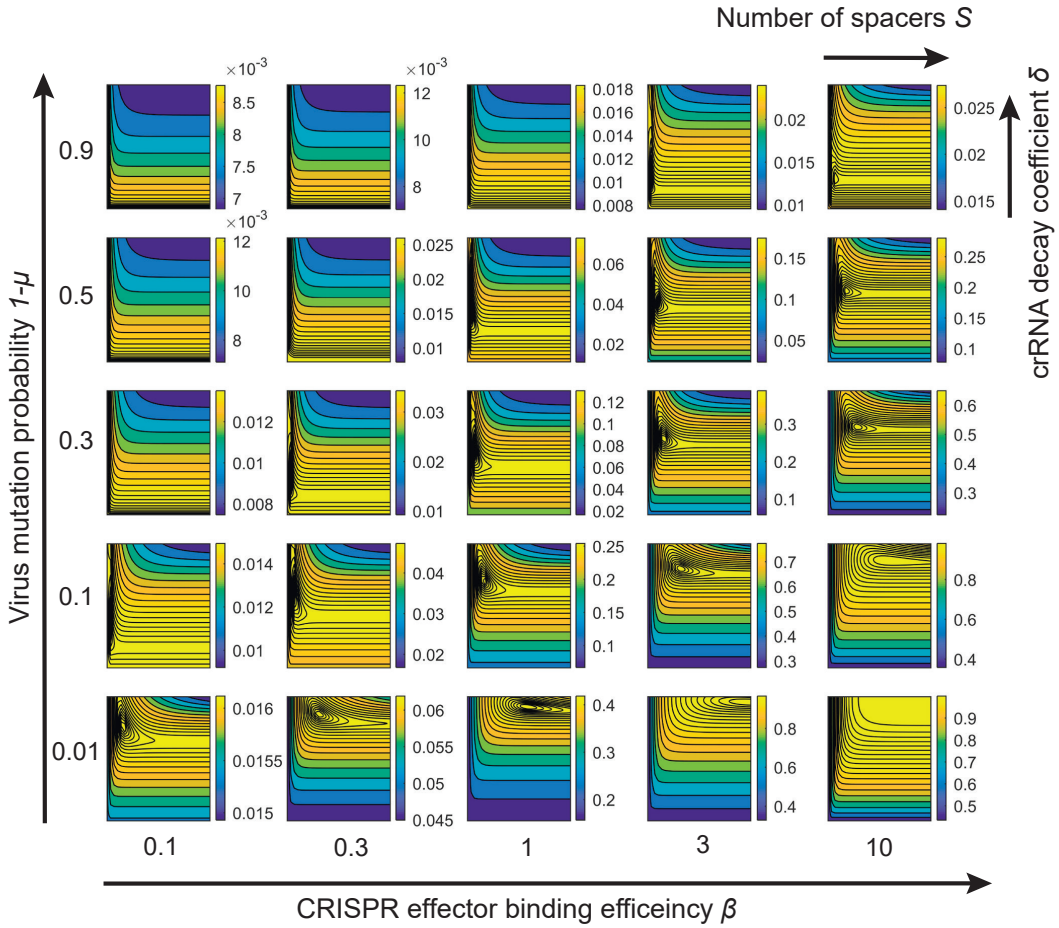

FIG. 4: **Effects of mutation rate and binding efficiency.** A set of 25 panels illustrating how the survival probability depends on S and δ for various values of protospacer mutation probability $1-\mu$ and binding efficiency of effectors β. The δ and S axes in each small panel have the same range as 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 $\chi = 2$ and $rNt = 5$.

of degradation of viral DNA should be high. This set of parameters favors arrays with more spacers. It implies a simple rule: the array can contain many spacers and be efficient or contain few spacers and be inefficient. In reality, the array composition could change on the timescale of viral infections, which may increase CRISPR interference efficiency. This, however, goes beyond the important assumption of our model of the static nature of the array and thus is beyond our present consideration. On the other hand, it shreds the light on the adaptive immunity as the only

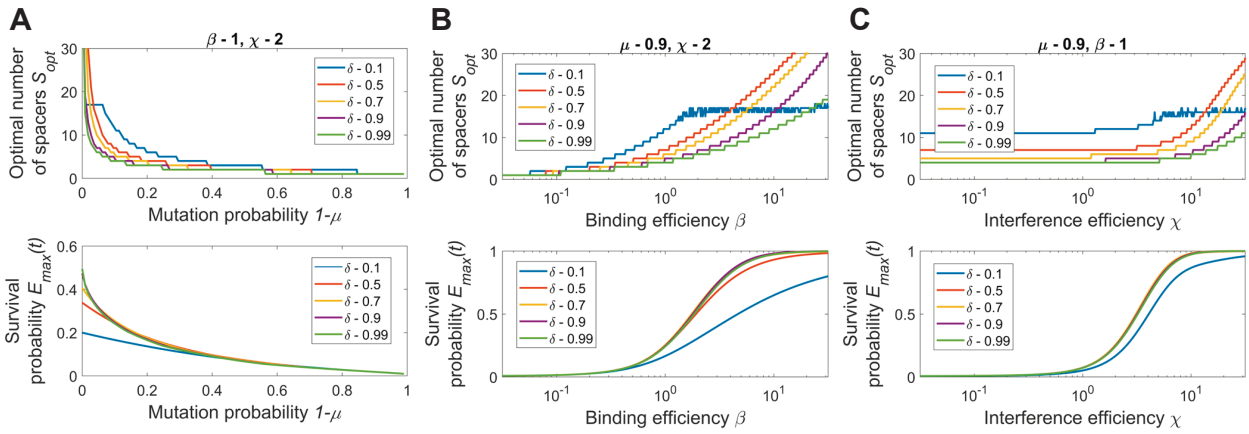

FIG. 5: **Effect of parameters on the optimal number of spacers and the maximal survival probability.** The optimal number of spacers and corresponding survival probability as functions of one of the array-unrelated parameters: (A) As function of mutation probability $1-\mu$, other parameters are $\beta = 1$ and $\chi = 2$. (B) As function of binding efficiency β, other parameters are $\mu = 0.9$ and $\chi = 2$. (C) As function of interference efficiency χ, other parameters $\mu = 0.9$ and $\beta = 1$. The average number of viral infections was $rNt = 5$ in all panels.

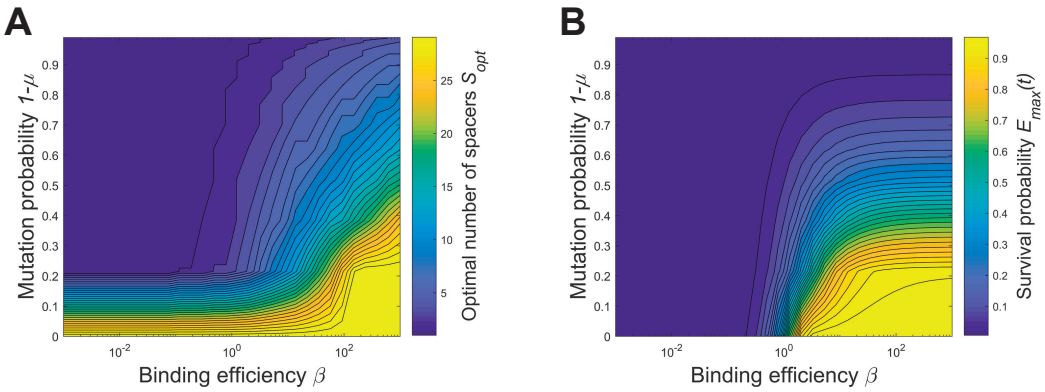

FIG. 6: **The optimal number of spacers and maximal cell survival probability.** The optimal number of spacers (A) and the maximal cell survival probability (B) are shown vs. a range of binding efficiencies β and mutation probabilities $1-\mu$ for $rNt = 5$ and $\chi = 2$.

efficient way of CRISPR-related defense in the viral environments with fast mutation rates.

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 $\exp(-rNtv_j)$ 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 v_i the fraction of the ith virus in the total number of viruses N so that $v_i = N_i/N$, 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,

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

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 equation (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 certain combination of spacers is

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

where v_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 v_1 and v_2 can be in any of the following four forms with corresponding probabilities: $P_{(1,1)} = v_1^2$, $P_{(1,2)} = P_{(2,1)} = v_1 v_2$, and $P_{(2,2)} = v_2^2$.

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.

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

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 $\chi = 2$, 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 certain δ and 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 additional viral species will affect the optimal S and δ more strongly. However, relating the model parameters to the experimental results [31], it is unlikely that the CRISPR efficiency in the multivirus environment 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

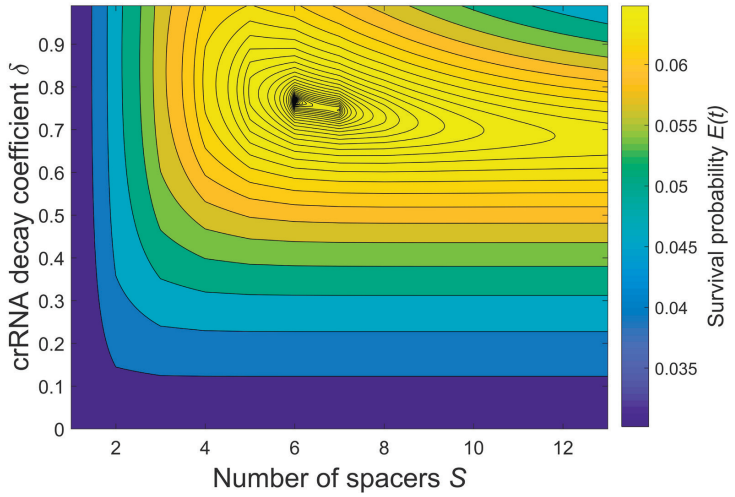

FIG. 7: **CRISPR performance for two virus species.** Plot of the survival probability $E(t)$ as a function of crRNA decay coefficient δ and the number of spacers S of a cell confronting two different viruses with equal population sizes, $\nu_1 = \nu_2 = 0.5$. The binding efficiency is $\beta = 1$ and the interference efficiency is $\chi = 2$. Viral mutation probability $1 - \mu$ is equal to 0.1 and $rNt = 5$.

minimum when the two 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.

DISCUSSIONS AND CONCLUSIONS

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. The interplay of these

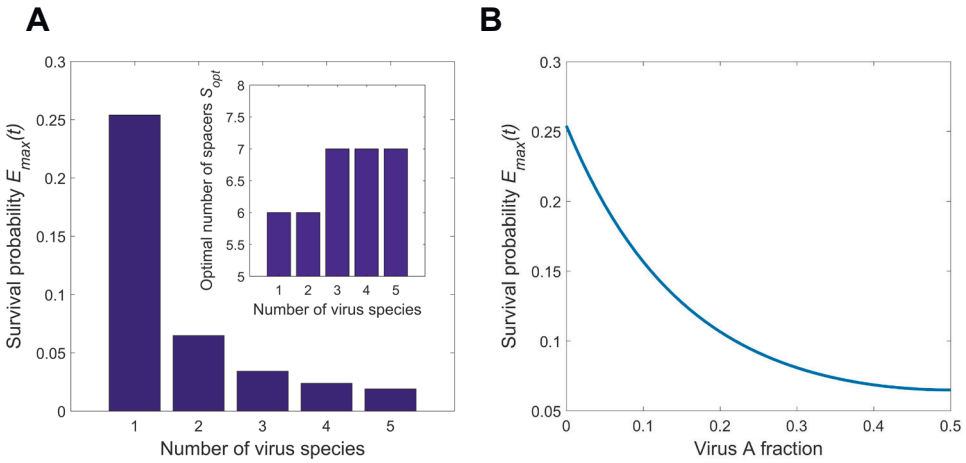

FIG. 8: **Survival probability vs diversity of the virus pool.** Plots of the optimized over δ and S cell survival probability and the number of spacers vs the number of viral species and the composition of a two-virus pool for $\beta = 1$, $\chi = 2$, $\mu = 0.9$ and $rNt = 5$. (A) Maximal survival 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, $\nu_1 = ... = \nu_n = 1/n$. (B) The maximal survival probability vs the relative abundance of one of the viruses in a two-virus pool.

forces leads to the optimum in the number of spacers per array, determined by the properties of the CRISPR system and viral environment in the following way: A better binding of the CRISPRS effectors to their targets and faster rate of target DNA degradation allow a bacterium 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. A few comments on applicability of our results and biological insights that can follow from them are in order:

Effects of dynamics and environment.

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, 32]. While in our analysis we assumed that the viral environment is constant (except for appearance of mutant protospacers), the actual viral dynamics may affect the

optimal lengths of CRISPR arrays. Changes in the environment could explain an increased efficiency of the shorter array in experimental condition comparing to the wild strain [33]. Moreover, the primed acquisition of spacers could happen on the timescale of a virus attack [34], which would invalidate our basic assumption of separation of timescales of viral attacks and spacer acquisition. These factors, not analyzed in our work, could affect the optimal number of CRISPR spacers and are subject to further analysis.

Comparison with existing results

Our results generally agree with the main findings of models existing in the field: We confirm that a higher diversity of viruses in the environment results in a dominance of viruses over the CRISPR system [22, 35]. 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 [35]. 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 [31]. Our analysis based on [31] resulted in the estimate of the CRISPR failure probability around 30% compared to 10^{-5} in [21].

Importance of hairpins.

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 [36, 37]. It is well known that hairpins have a potential to pause or terminate transcription elongation [38, 39]. 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.

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 [40]. 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 [41], the cost of maintenance of cas genes [19], self-immunity [42] 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 length of the spacer array [43–45]. 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 length of the 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 [46].

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. Conversely, less favorable conditions shift the optimum to arrays with fewer spacers.
- Arrays containing only a few (less than 10) spacers offer significantly less protection from viral attacks.
- 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.

ACKNOWLEDGMENTS

Y.I. was supported by FONDECYT (Chile) grant 1151524 and by SkolTech during his stay there.

[1] K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin, Biology direct **1**, 7 (2006), ISSN 1745-6150, URL http://www.biology-direct.com/content/1/1/7.

[2] A. Bolotin, B. Quinquis, A. Sorokin, and S. Dusko Ehrlich, Microbiology **151**, 2551 (2005), ISSN 13500872.

[3] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. a. Romero, and P. Horvath, Science **315**, 1709 (2007), ISSN 0036-8075, URL http://www.sciencemag.org/cgi/doi/10.1126/science.1138140.

[4] K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, et al., Nature reviews. Microbiology **9**, 467 (2011), ISSN 1740-1534, URL http://dx.doi.org/10.1038/nrmicro2577.

[5] K. S. Makarova, Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. J. Brouns, E. Charpentier, D. H. Haft, et al., Nature Reviews Microbiology **13**, 722 (2015), ISSN 1740-1526, URL http://www.nature.com/doifinder/10.1038/nrmicro3569.

[6] K. R. Hargreaves, C. O. Flores, T. D. Lawley, and M. R. J. Clokie, mBio **5**, e01045 (2014), ISSN 2150-7511, URL http://mbio.asm.org/cgi/doi/10.1128/mBio.01045-13.

[7] G. C. McGhee and G. W. Sundin, PLoS ONE **7** (2012), ISSN 19326203.

[8] A. van Belkum, L. B. Soriaga, M. C. LaFave, S. Akella, J.-b. Veyrieras, E. M. Barbu, D. Shortridge, B. Blanc, G. Hannum, G. Zambardi, et al., mBio **6**, 1 (2015), ISSN 2150-7511, URL http://www.ncbi.nlm.nih.gov/pubmed/26604259.

[9] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, Nucleic Acids Research **41**, 4360 (2013), ISSN 03051048.

[10] Y. Agari, K. Sakamoto, M. Tamakoshi, T. Oshima, S. Kuramitsu, and A. Shinkai, Journal of Molecular Biology **395**, 270 (2010), ISSN 00222836, URL http://dx.doi.org/10.1016/j.jmb.2009.10.057.

[11] D. Rath, L. Amlinger, A. Rath, and M. Lundgren, Biochimie **117**, 119 (2015), ISSN 61831638, URL http://dx.doi.org/10.1016/j.biochi.2015.03.025.

[12] C. Díez-Villaseñor, C. Almendros, J. García-Martínez, and F. J. M. Mojica, Microbiology **156**, 1351 (2010), ISSN 13500872.

[13] P. Horvath, D. A. Romero, A. C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou, Journal of Bacteriology **190**, 1401 (2008), ISSN 00219193.

[14] I. Grissa, G. Vergnaud, and C. Pourcel, BMC bioinformatics **8**, 172 (2007), ISSN 14712105.

[15] C. Hale, K. Kleppe, R. M. Terns, and M. P. Terns, RNA (New York, N.Y.) **14**, 2572 (2008), ISSN 1469-9001, URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2590957{\&}tool=pmcentrez{\&}rendertype=abstract.

[16] B. R. Levin, S. Moineau, M. Bushman, and R. Barrangou, PLoS Genetics **9** (2013), ISSN 15537390.

[17] L. A. Marraffini and E. J. Sontheimer, Science **322**, 1843 (2008), ISSN 0036-8075, NIHMS150003, URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{\&}id=19095942{\&}retmode=ref{\&}cmd=prlinks{\%}5Cnpapers2:

//publication/doi/10.1126/science.1165771.

[18] J. Bondy-Denomy and A. R. Davidson, Trends in Microbiology **22**, 218 (2014), ISSN 18784380, URL http://dx.doi.org/10.1016/j.tim.2014.01.007.

[19] P. F. Vale, G. Lafforgue, F. Gatchitch, R. Gardan, S. Moineau, and S. Gandon, Proceedings of the Royal Society B: Biological Sciences **282**, 20151270 (2015), ISSN 0962-8452, URL http://dx.doi.org/10.1098/rspb.2015.1270{\%}5Cnhttp://rspb.royalsocietypublishing.org.http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2015.1270.

[20] U. Gophna, D. M. Kristensen, Y. I. Wolf, O. Popa, C. Drevet, and E. V. Koonin, The ISME journal **9**, 2021 (2015), ISSN 1751-7370, URL http://dx.doi.org/10.1038/ismej.2015.20.

[21] L. M. Childs, N. L. Held, M. J. Young, R. J. Whitaker, and J. S. Weitz, Evolution **66**, 2015 (2012), ISSN 00143820, URL http://doi.wiley.com/10.1111/j.1558-5646.2012.01595.x.

[22] J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Journal of Bacteriology **195**, 3834 (2013), ISSN 00219193.

[23] S. Bradde, M. Vucelja, T. Tesileanu, and V. Balasubramanian, PLOS Computational Biology **13**, e1005486 (2017), ISSN 1553-7358, 1510.06082, URL http://arxiv.org/abs/1510.06082http://dx.plos.org/10.1371/journal.pcbi.1005486.

[24] C. Díez-Villaseñor, N. M. Guzmán, C. Almendros, J. García-Martínez, and F. J. Mojica, RNA Biology **10**, 792 (2013), ISSN 1547-6286, URL http://www.tandfonline.com/doi/full/10.4161/rna.24023.

[25] S. A. Jackson, R. E. McKenzie, R. D. Fagerlund, S. N. Kieper, P. C. Fineran, and S. J. J. Brouns, Science **356**, eaal5056 (2017), ISSN 0036-8075, URL http://www.sciencemag.org/lookup/doi/10.1126/science.aal5056.

[26] E. Semenova, M. M. Jore, K. a. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, Proceedings of the National Academy of Sciences of the United States of America **108**, 10098 (2011), ISSN 0027-8424.

[27] S. Fischer, L. K. Maier, B. Stoll, J. Brendel, E. Fischer, F. Pfeiffer, M. Dyall-Smith, and A. Marchfelder, Journal of Biological Chemistry **287**, 33351 (2012), ISSN 00219258.

[28] S. Shah, S. Erdmann, F. Mojica, and R. Garrett, RNA biology **10**, 891 (2013), ISSN 1547-6286, URL http://www.landesbioscience.com/journals/rnabiology/

2012RNABIOL0169R.pdf.

[29] Ü. Pul, R. Wurm, Z. Arslan, R. Geißen, N. Hofmann, and R. Wagner, Molecular Microbiology **75**, 1495 (2010), ISSN 13652958.

[30] J. Zoephel and L. Randau, Biochemical Society Transactions **41**, 1459 (2013), ISSN 0300-5127, URL http://biochemsoctrans.org/lookup/doi/10.1042/BST20130129.

[31] A. Strotskaya, E. Savitskaya, A. Metlitskaya, N. Morozova, K. A. Datsenko, E. Semenova, and K. Severinov, Nucleic Acids Research p. gkx042 (2017), ISSN 0305-1048, URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkx042.

[32] F. S. Berezovskaya, Y. I. Wolf, E. V. Koonin, and G. P. Karev, Biology direct **9**, 13 (2014), ISSN 1745-6150, URL http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-9-13http://www.ncbi.nlm.nih.gov/pubmed/24986220{\%}5Cnhttp://www.biologydirect.com/content/pdf/1745-6150-9-13.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/24986220http://www.pubmedcentral.nih.gov/.

[33] C. Rao, D. Chin, and A. W. Ensminger, bioRxiv (2017).

[34] E. Semenova, E. Savitskaya, O. Musharova, A. Strotskaya, D. Vorontsova, K. A. Datsenko, M. D. Logacheva, and K. Severinov, Proceedings of the National Academy of Sciences of the United States of America **113**, 7626 (2016), ISSN 1091-6490 (Electronic).

[35] A. D. Weinberger, Y. I. Wolf, A. E. Lobkovsky, M. S. Gilmore, and E. V. Koonin, mBio **3**, 1 (2012), ISSN 21507511.

[36] V. Kunin, R. Sorek, and P. Hugenholtz, Genome biology **8**, R61 (2007), ISSN 1474-760X, URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-4-r61.

[37] S. J. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. H. Slijkhuis, A. P. L. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. Van Der Oost, Cancer Epidemiology Biomarkers and Prevention **2**, 531 (1993), ISSN 10559965, 20.

[38] K. S. Wilson and P. H. V. Hippel, Biochemistry **92**, 8793 (1995), ISSN 0027-8424.

[39] P. J. Farnham and T. Platt, Nucleic Acids Research **9**, 563 (1981), ISSN 03051048.

[40] P. Han and M. W. Deem, Journal of The Royal Society Interface **14**, 20160905 (2017), ISSN 1742-5689, URL http://rsif.royalsocietypublishing.org/lookup/doi/10.1098/rsif.2016.0905.

[41] C. H. Kuo and H. Ochman, Genome Biology and Evolution **1**, 145 (2010), ISSN 1759-6653, URL https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evp016.

[42] R. B. Vercoe, J. T. Chang, R. L. Dy, C. Taylor, T. Gristwood, J. S. Clulow, C. Richter, R. Przybilski, A. R. Pitman, and P. C. Fineran, PLoS Genetics **9** (2013), ISSN 15537390.

[43] Y. Wei, R. M. Terns, and M. P. Terns, Genes & Development **29**, 356 (2015), ISSN 0890-9369, URL http://genesdev.cshlp.org/lookup/doi/10.1101/gad.257550.114.

[44] A. Levy, M. G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, and R. Sorek, Nature **520**, 505 (2015), ISSN 0028-0836, URL http://www.nature.com/doifinder/10.1038/nature14302.

[45] I. Yosef, M. G. Goren, and U. Qimron, Nucleic Acids Research **40**, 5569 (2012), ISSN 03051048.

[46] W. Jiang, I. Maniv, F. Arain, Y. Wang, B. R. Levin, and L. A. Marraffini, PLoS Genetics **9**, e1003844 (2013), ISSN 1553-7404, URL http://dx.plos.org/10.1371/journal.pgen.1003844.