SKOLKOVO INSTITUTE OF SCIENCE AND TECHNOLOGY

USING MATHEMATICAL MODELING TO UNDERSTAND PROKARYOTIC ADAPTIVE IMMUNITY

Doctoral Thesis

by

ALEXANDER MARTYNOV

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor

Professor, Konstantin Severinov

Co-supervisor

Dr. Jaroslav Ispolatov

Moscow - 2018

Abstract

Alexander Martynov

Using mathematical modeling to understand prokaryotic adaptive immunity

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR associated protein) is a prokaryotic adaptive immune system. It is an anti-viral tool which recognizes and destroys viruses and other foreign genetic elements including plasmids, which DNA is recorded in CRISPR array. Being a recent discovery, the CRISPR-Cas system holds a broad range of unknown mechanisms, unique and not fully understood features and paradoxes. In this work, we analyzed various aspects of CRISPR-Cas functioning through mathematical modeling. In the first part of this work, 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. In the second part of this work, we focused on the interaction of CRISPR-Cas systems and plasmids. We show that some plasmids carrying a protospacer matching crRNA spacer can be stably maintained in cells at conditions of ongoing CRISPR interference. We proposed a model that explains this observation based on opposing kinetics of plasmid replication and CRISPR-Cas interference. If plasmid number in a cell stochastically increases despite the average dominance of interference over replication, a "plasmid stability window" may be reached where plasmid duplication rate is higher than the interference rate, leading to stable plasmid maintenance. Stochastic simulations reveal that under pressure from CRISPR-Cas, the initially uniform plasmid copy number distribution in cell population becomes bimodal: one fraction of cells loses plasmids and becomes plasmid-free, while another fraction of the cells keeps the plasmids. We conducted a series of experiments with *Escherichia coli* cells with activated I-E CRISPR-Cas confirming the bimodal plasmid distribution predicted by the model.

Publications

Publications on the topic of the dissertation:

- Optimal number of spacers in CRISPR arrays. **Martynov A**, Severinov K, Ispolatov I. PLoS Comput Biol. 2017 Dec 18;13(12):e1005891. doi: 10.1371/journal.pcbi.1005891. eCollection 2017 Dec.

Publications on other topics during the doctoral program:

- **Martynov A.G.**, Elpidina E.N., Perkin L. and Oppert B., Functional analysis of C1 family cysteine peptidases in the larval gut of Tenebrio molitor and Tribolium castaneum. BMC Genomics. 2015 Feb 14;16:75 doi: 10.1186/s12864-015-1306-x

- Savir Y.*, **Martynov A.***, Springer M. (2017) Achieving global perfect homeostasis through transporter regulation. PLoS Comput Biol 13(4): e1005458. doi: 10.1371/journal.pcbi.1005458

- Oppert B., Perkin L., **Martynov A.G.**, Elpidina E.N.. Cross-species comparison of the gut: Differential gene expression sheds light on biological differences in closely related tenebrionids. J Insect Physiol. 2017 Mar 28. doi: 10.1016/j.jinsphys.2017.03.010

- Schoville SD, Chen YH, Andersson MN, Benoit JB, Bhandari A, Bowsher JH, Brevik K, Cappelle K, Chen MM, Childers AK, Childers C, Christiaens O, Clements J, Didion EM, Elpidina EN, Engsontia P, Friedrich M, García-Robles I, Gibbs RA, Goswami C, Grapputo A, Gruden K, Grynberg M, Henrissat B, Jennings EC, Jones JW, Kalsi M, Khan SA,

Kumar A, Li F, Lombard V, Ma X, **Martynov A**, Miller NJ, Mitchell RF, Munoz-Torres M, Muszewska A, Oppert B, Palli SR, Panfilio KA, Pauchet Y, Perkin LC, Petek M, Poelchau MF, Record É, Rinehart JP, Robertson HM, Rosendale AJ, Ruiz-Arroyo VM, Smagghe G, Szendrei Z, Thomas GWC, Torson AS, Vargas Jentzsch IM, Weirauch MT, Yates AD, Yocum GD, Yoon JS, Richards S. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). Sci Rep. 2018 Jan 31;8(1):1931. doi: 10.1038/s41598-018-20154-1

* Co-first authorship

Acknowledgements

I am grateful to all my supervisors, Konstantin Severinov who supported me during my PhD journey and taught me how to conduct the research and manage myself as a scientist, Jaroslav Ispolatov who taught me most of the concepts of mathematical modeling and helped me to develop a theoretician intuition and Michael Springer who taught me lots of useful techniques for writing scientific papers.

I would like to thank all Konstantin Severinov lab members who helped me during this research. In particular, I would specially thank Viktor Mamontov, Ekaterina Semenova and Alexandra Strotskaya who made an invaluable contribution helping in the design and execution of experimental parts of the work.

Last but not least, I would also like to thank all my friends and family members that supported me on this journey.

Contents

List of Figures

List of Tables

Dedicated to my mom for all
the effort she had spent to
teach me.

Chapter 1

Literature review

1.1 CRISPR-Cas mechanism of action

1.1.1 Main components of CRISPR-Cas system and their roles

CRISPR (Clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated proteins) is a system of prokaryotic adaptive immunity [1, 2]. CRISPR-Cas system consists of two main components: CRISPR array that serves as a recognition library and Cas-proteins that are effectors of the CRISPR-Cas immunity.

CRISPR repeat-spacer array is a chain of identical repeats that separate unique spacers - the DNA fragments that match foreign DNA and thus can be used to compare with and recognize the foreign DNA. Being transcribed, CRISPR spacers form CRISPR RNA (crRNA) can interfere with corresponding sequences of foreign DNA or RNA, by guiding Cas-proteins to it [1]. CRISPR array functions as a memory of a CRISPR-Cas system recording previous infections and as a recognition key that guides Cas proteins.

Cas protein genes are typically organized as compact clusters localized nearby the CRISPR array that is transcribed together. Cas proteins (and corresponding *cas* genes) composition varies from species to species but most of

the CRISPR-Cas systems consist of three different protein types [3, 4, 5]. Immunity acquisition proteins Cas1-Cas2 are responsible for adaptation - capture and storing of new spacers in CRISPR array [6]. Recognition and nuclease Cas proteins are responsible for the elimination of target DNA. They include Cas3 and the Cascade complex in type I CRISPR-Cas systems or Cas9 in type II CRISPR-Cas systems. Other supporting and regulatory proteins help during the establishment of the effector complex (Cas proteins and crRNA).

1.1.2 Discovery and classification of CRISPR-Cas systems

CRISPR-Cas systems originally were discovered in late 80-s, in a report on unusual genomic repeats in Escherichia coli [7]. However, its possible function as an immunity system was not shown until 2005 when three different groups demonstrated that some spacers of CRISPR array correspond to viral genomes [2, 8, 9]. During the first steps of investigations of CRISPR-Cas there were no true systematics and all *cas* genes and novel systems were named inconsistently. Nowadays all CRISPR-Cas systems are classified based on the Cas protein composition. Originally all found CRISPR-Cas systems were separated into three subtypes: CRISPR-Cas type I, type II and type III [3]. Each of those had distinct Cas protein composition and differences in mechanisms of actions.

With the advances of computational techniques that allowed automated CRISPR-Cas systems discovery [10] more distinct types of CRISPR-Cas systems were found. This allowed identification of rare or unique and structurally different systems and revealed the true diversity of CRISPR-Cas systems. As a result, an updated CRISPR-Cas systems classification was proposed [4]; it divided all CRISPR-Cas systems into two classes which are further sub-divided into types (see fig. 1.1).

Class 1 CRISPR-Cas systems rely on large effector complexes made up of multiple Cas proteins. Class 1 CRISPR-Cas systems are sub-categorized into type I, type III and type IV, based on the Cas-protein composition. All these three types seemingly have evolved from the same origin [11, 12]. These systems can coexist in the same organism and potentially can be compatible with each other and share the same CRISPR arrays. Yet, systems of different types differ in their mechanism of action (see further sections).

Type I CRISPR-Cas systems are characterized by the presence of the Cas3 signature protein which holds both helicase and nuclease activities [13]. Also type I systems are distinguished by a set of *cas* genes that code for proteins forming a CRISPR-Associated Complex for Antiviral Defense (Cascade) complex: they typically include *cas5*, *cas7* and some variant of *cas8*. Also type CRISPR-Cas loci typically include *cas1*, *cas2* and, some also *cas4* genes that encode proteins involved in adaptation. Type III CRISPR-Cas is distinguished by the Cas10 signature protein. It is also remarkable that most type III CRISPR-*cas* loci do not contain *cas1* and *cas2* genes. Moreover, there are typically no reported so far CRISPR arrays associated with type III *cas* genes. A relatively novel type IV CRISPR-Cas is a rare and rudimentary system that was predicted in several bacteria. It is characterized by the signature gene *scf1* and a relatively small number of other *cas* genes: alongside with mentioned *scf1* type IV systems typically include only *cas5* and *cas7* [4, 12].

Class 2 CRISPR-Cas systems are distinguished by a large single protein that, when bound to crRNA, performs all functions of the effector complex [4, 5]. These large effector proteins typically consist of two lobes: recognition lobe (REC) that holds crRNA and is responsible for target recognition, and nuclease lobe (NUC) that is responsible for target degradation. Class 2 CRISPR-Cas systems are also subdivided into three types: type II, type V, and type VI [5]. Class 2 types are structurally dissimilar from each other and seemingly have undergone convergent evolution. While nuclease lobes of

type II and type V CRISPR-Cas show some conserved motifs, their recognition lobes show no similarity. Type VI CRIPSR-Cas shows no evolutionary similarity with other types and must have evolved independently from a completely different set of ancestor proteins. [3, 4, 5].

Type II CRISPR-Cas is characterized by the presence of the *cas9* gene that encodes one of the most studied effectors [14]. Also in type II CRISPR-Cas loci there is a typical presence of Cas1-Cas2 proteins and a tracrRNA, which is important in crRNA maturation and targeting (see further sections). Type V CRISPR-Cas systems include one of the Cas12 effector variants, which, as Cas9, contains a RuvC nuclease domain. However, its recognition lobe has a different structure that distinguishes type V systems from type II systems. Type VI CRISPR-Cas is the most novel system that relies on Cas13 effectors (previously named C2c2). It is distinguished by the presence of two HEPN domains, present in many RNA binding proteins [15]

Overall it is fair to say that most probably the current CRISPR-Cas classification is incomplete and temporary and future discoveries might extend, shape or restructure the current view of CRISPR-Cas, their evolution and interrelationships to each other.

1.1.3 Mechanism of interference

As described above, a CRISPR defense mechanism relies on the degradation of foreign DNA or RNA by Cas proteins that are guided to targets by crRNA transcribed from the CRISPR array. This process is called CRISPR-interference [1]. It is regulated on multiple levels in order to ensure efficient anti-viral defense and avoidance of self-immunity. Different types of CRISPR-Cas systems have different sets of interfering Cas proteins that often perform the similar function but are very different structurally [5]. These complexes seem to undergo convergent evolution and while conceptually the

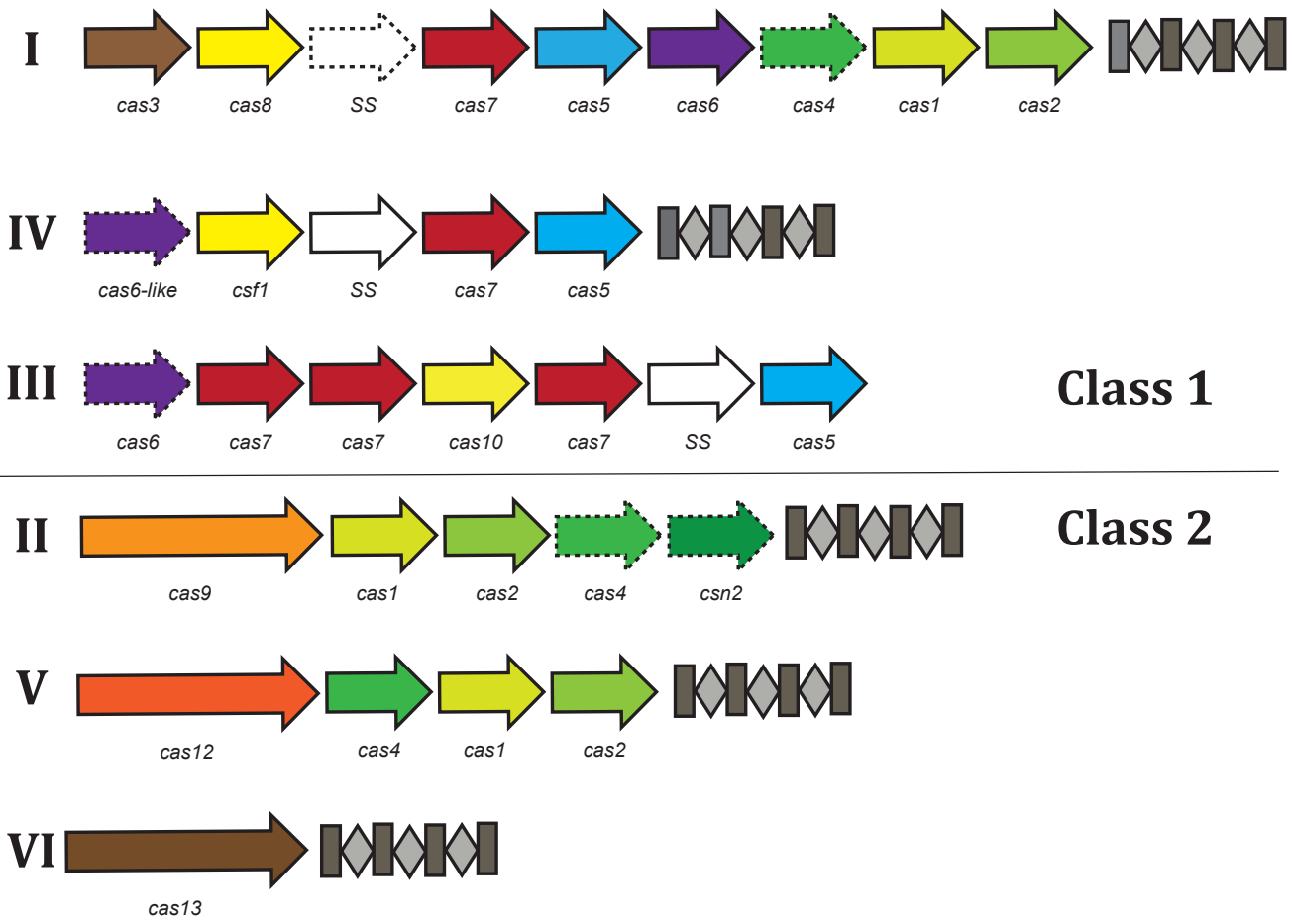

FIGURE 1.1: **Classification of CRISPR-Cas systems.** Schematics of typical *cas* genes organization of CRISPR-Cas systems of different types based on [5]. Arrows represent *cas* genes: solid arrows correspond to obligatory genes that are present in all studied organisms while dashed arrows correspond to "optional" genes that are present only in some CRISPR-Cas systems of a given type. Gray boxes correspond to repeats and light gray rhombi correspond to spacers forming a CRISPR array. SS stands for small subunit. Gene order for type I is presented according to typical arrangements in CRISPR-Cas systems from *Escherichia coli* K12 and *Bacillus halodurans*; for type II - according to *Legionella pneumophila str. Paris* and *Streptococcus thermophilus*; for type III - according to *Methanothermobacter thermautotrophicus* and *Staphylococcus epidermidis*; for type IV - according to *Thioalkalivibrio sp. K90mix* and *Rhodococcus jostii RHA1*; for type V - according to *Francisella cf. novicida Fx1*; for type VI - according to *Fusobacterium prefoetens*.

process of CRISPR interference is similar within CRISPR-Cas systems of different types most of the mechanistic details significantly differ (see fig. 1.2).

The first step of CRISPR-Cas interference is the formation of the effector complex that can target foreign DNA. This includes transcription of *cas* genes, translation of effector polypeptides and charging effector complexes with crRNA. CRISPR array is transcribed as a long precursor CRISPR RNA (pre-crRNA) [1, 16, 17]. Then it is being processed into short mature crRNA, each corresponding to an individual spacer. The mechanism of this processing varies between different types of CRISPR-Cas systems [18]. In type I CRISPR-Cas, processing of crRNA is performed by one of the subunits of the Cascade effector complex - Cas6 [19]. After cleavage crRNA stays bound to the Cascade complex. In type II CRISPR-Cas systems there is a single large Cas9 protein that performs all interference activity. In this case, the processing of crRNA is rather complex - it requires additional small RNA - the transencoded crRNA (tracrRNA). This tracrRNA has a specific structure that can form a hairpin that allows its binding to Cas9 and also has a region complementary to repeat sequences in pre-crRNA. Mediated by Cas9, tracrRNA forms a double-stranded RNA duplex with pre-crRNA that is then processed by RNase III encoded outside the CRISPR-Cas locus [20]. In type III CRISPR-Cas systems pre-crRNA is also processed by Cas6 endoribonuclease which, however, is not a part of the effector complex [16]. Type III Cas6 produces intermediate crRNAs that are being further trimmed at the 3′ end by other distinct but yet unidentified factors [21]. Mature crRNA is transferred to effector Cas10–Csm or Cas10–Cmr complex for type III-A or III-B CRISPR-Cas systems, respectively [22].

When CRISPR effector complexes are formed the interference - recognition of foreign DNA and its degradation can occur. In type I systems recognition is performed by the Cascade [17] - a large complex that is formed by different Cas proteins. In order to achieve successful target recognition type I effectors require the so-called protospacer adjacent motif (PAM) [23]. It is a specific sequence located next to a protospacer (and therefore outside of spacer complementarity region) that is recognized by the Cse1 (CasA) subunit of Cascade [24] and allows the Cas3 nuclease to bind [25]. The PAM sequence is extremely important for self- versus non-self recognition. Since all crRNAs are transcribed from CRISPR array, they sequence obviously matches their origin in CRISPR array. However, since PAM is absent from the part of repeat and the spacer-repeat junction distinguishing of foreign DNA from the CRISPR array becomes possible. The first 8 nucleotides of the spacer form the so-called "seed" region, which is crucial for target recognition. Mutations in the seed region that introduce single mismatches with the target almost completely abolish interference while changes in other parts of do not lead to loss of recognition and could still result in successful interference (but sometimes at lower efficiency) [23]. In effector complexes, the seed region is typically specifically pre-ordered by Cas proteins which establish a specific conformation in order to enhance protospacer binding [26]. The seed region presumably is the first to be matched between spacer and protospacer, making complex formation thermodynamically favorable. Upon a successful spacer-protospacer match Cascade complex recruits Cas3 – an ATP-dependent helicase and single-strand DNA nuclease which initiates target DNA degradation [13]. Cas3 slides over the target DNA introducing periodic single-strand breaks, which allows the foreign DNA to be targeted by other cell nucleases.

In type II systems Cas9 combines both the target recognition and nuclease functions [27, 28]. Type II systems also require PAM for target protospacer recognition [8, 27]. PAM is recognized by a specific domain of Cas9 [29].

Foreign DNA degradation, in this case, occurs through the introduction of a double-strand break in the target [30], catalyzed by two domains of Cas9: HNH and RuvC [31]. Each of these domains cut their own strand of target DNA [31].

In type III CRISPR-Cas systems the recognition occurs by a completely different mechanism. The effectors are Cas10–Csm or Cas10–Cmr complexes for type III-A or III-B systems, respectively, they contain multiple subunits and are similar in function to the Cascade complex of type I systems [21]. Self- versus non-self differentiation in the case of type III systems is PAM-independent. crRNAs processed from pre-crRNA contain a fragment of the repeat - so-called crRNA tag [16, 32]. This crRNA tag is required in order to probe target sequence and only sequences that do not possess full match to the tag sequence are destroyed [32]. CRISPR array or its transcripts fully match crRNA and remain unrecognized, preventing autoimmunity. Another distinct feature of type III CRISPR-Cas complexes is that they target both RNA and DNA [33, 34]. Subunits Csm3 and Cmr4 cleave single-stranded RNAs complementary to crRNA into fragments of discrete lengths [33, 35]; the Csm1 and Cmr2 subunits cleave DNA introducing single-strand breaks [33, 36]. The latter activity is dependent on complementary RNA recognition.

There are several reports that neither Cas proteins nor crRNA *in vivo* are produced constitutively [37, 38]. Instead, their production is regulated by various external events. In *Escherichia coli* the type I CRISPR-Cas is repressed by a global transcription repressor – DNA binding protein H-NS [37]. H-NS binding could be countered by transcription factor LeuO, an H-NS antagonist. LeuO overexpression activates E. coli CRISPR response, promoting both interference and new spacer acquisition[38, 39]. LeuO expression can be triggered by various cellular stress conditions including cell membrane penetration by viruses [40].

Other newly found types of CRISPR-Cas systems, such as type IV, V

and VI are yet largely understudied [5]. Given their structural diversity, we might expect to find other mechanisms of protospacer recognition, self- versus non-self distinction and regulation. For instance, newly discovered type VI CRISPR-Cas systems effector complex contains RNAse Cas13 that target only RNA [10, 41] and must be controlled differently than DNA targeting systems. Overall, CRISPR-Cas systems can be viewed as a diverse group that only share a similar molecular memory organization but are quite distinct in mechanisms of action/use of this memory.

1.1.4 Mechanism of immunization

CRISPR interference relies on an existing database of spacers. Thus, initially, the CRISPR-Cas needs to gather immunity – undergo an immunization process, which is also called adaptation. Generally, this process includes recognition of foreign DNA, a capture of protospacer DNA fragments and their integration into the CRISPR array as new spacers. Adaptation could happen through two distinct mechanisms: naive adaptation - acquisition of a completely new spacer and primed adaptation - one that depends on recognition of foreign DNA by a pre-existing spacer.

Naive adaptation has been first shown nearly a decade ago and has been extensively studied since then. While the main mechanism has been revealed for type I CRISPR-Cas systems most details remain unclear or being debated. The naive adaptation consists of three stages: source DNA preparation, spacer acquisition, and spacer integration [42] (fig. 1.3).

The first step of naive adaptation is the formation of DNA fragments suitable for integration. It is currently believed that one way these fragments are produced is through the activity of the RecBCD exonuclease complex [42, 43]. RecBCD is a part of prokaryotic double-strand break repair system [44]. It recognizes the ends of double-stranded breaks in DNA, unwinds DNA from

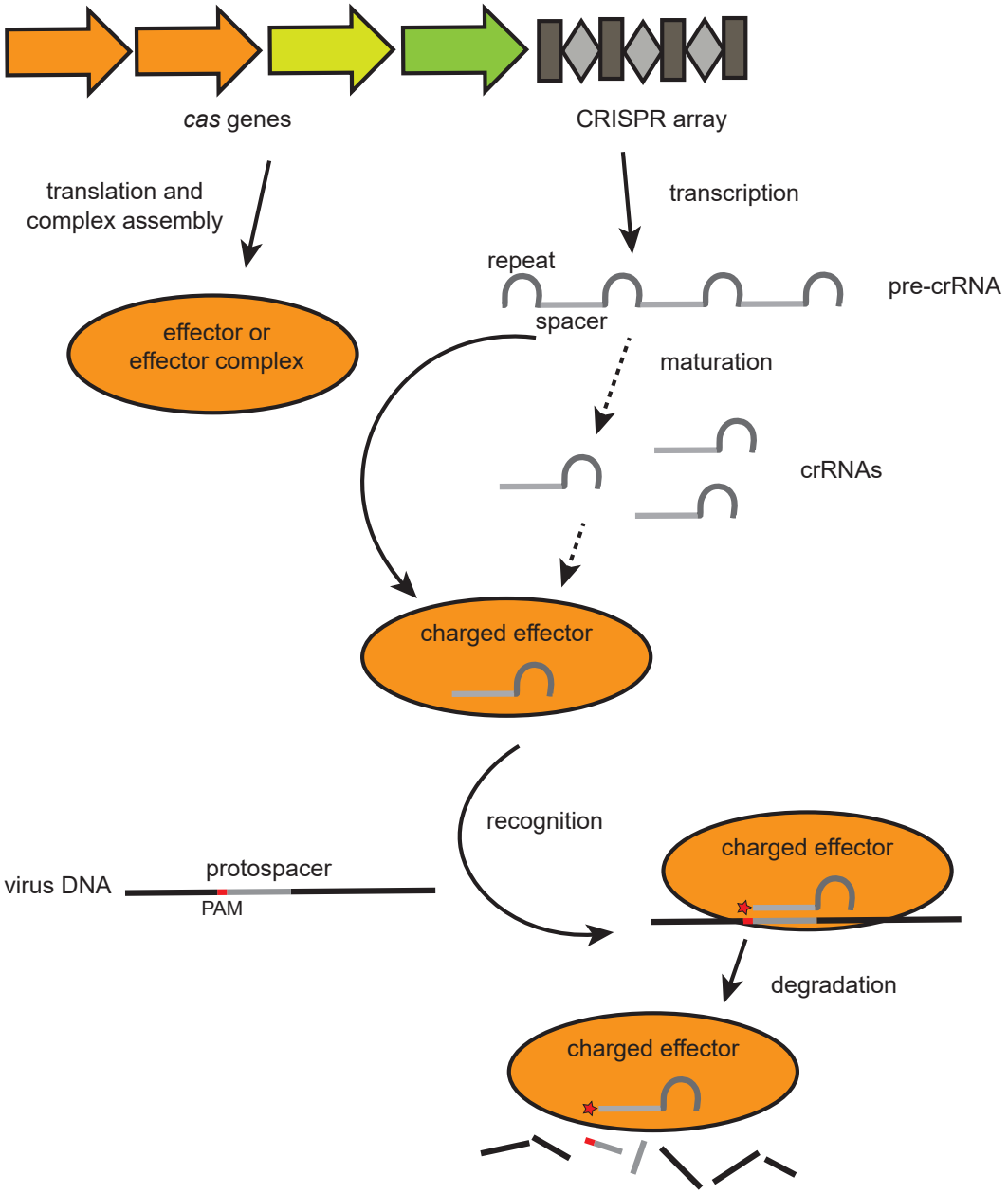

FIGURE 1.2: **Scheme of CRISPR-Cas interference process.** The scheme presented is closest to the mechanism of type I CRISPR-Cas systems. Effectors (orange) after being translated are charged with crRNAs which each consist of a spacer segment (light gray) and repeat hairpin (dark gray). In order for pre-crRNA to maturate it is cleaved by the subunits of the effector complex (solid curved arrow) as in type I CRISPR-Cas systems or by dedicated standalone RNAses and then bound by effectors (dashed arrows) as in type III systems. The viral protospacer (gray) and PAM (red) is recognized by crRNA spacer and PAM recognition subunit or domain (red star) of the effector and viral DNA is degraded.

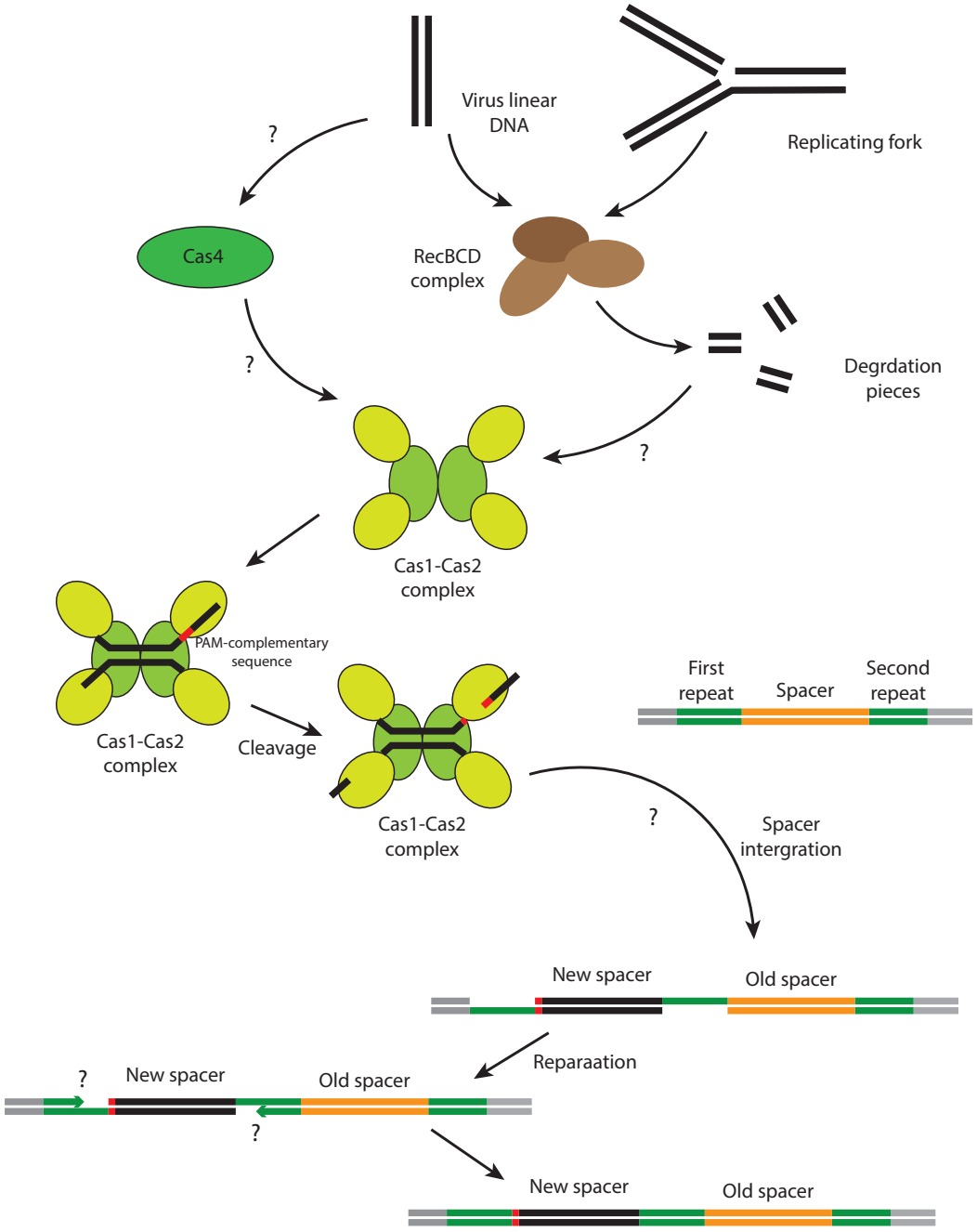

FIGURE 1.3: **Scheme of CRISPR-Cas naive adaptation.** Virus, plasmid or self DNA is degraded by RecBCD complex (three brown globes) and degradation pieces are uptaken by the Cas1-Cas2 complex. Alternatively, Cas4 can obtain viral DNA and transfer it to Cas1-Cas2 through an unknown mechanism. Cas2 subunit of the complex is shown by darker and Cas1 - by lighter green color. The new spacer is shown in black with the red part corresponding to PAM and PAM-complementary sequence, which is further trimmed by Cas1-Cas2 leaving only guanine and cytosine in PAM and PAM-complementary sequence, respectively. The new spacer is integrated splitting two chains of the first repeat (green). Then the remaining chain is being repaired (shown by green arrows) by unknown cellular polymerase and ligase.

the 3′ end [45] and moves along the DNA strand executing periodic single-stranded breaks. The produced DNA fragments might be used as a source of spacers by the Cas1-Cas2 complex [43]. The specificity of RecBCD makes it a good tool to specifically target foreign DNA for RecBCD can target genome ends of double-stranded DNA viruses such as the lambda phage [46]. Also, RecBCD could operate together with type II restriction enzymes that cleave foreign DNA into linear fragments by introducing double-stranded breaks [47]. Moreover, RecBCD complex also targets frequently replicating DNA such as that of plasmids and viruses [43], since it is recruited by the stalled or collapsed replication forks which are also a source of double-strand DNA ends [48, 49]. The problem of self-DNA degradation by RecBCD and assuring specificity towards foreign DNA as a source of new spacers may be resolved on two levels. On one hand, RecBCD degradation is inhibited by Chi sites [50, 51] which, when present, are highly overrepresented in the host genome compared to viral genomes [52]. This should limit RecBCD-mediated degradation to the nearest Chi-site in host DNA and shall not allow RecBCD to do much harm and also generate self-targeting spacer precursors. Further, more active replication leads to a bias towards the acquisition of spacers towards foreign DNA [43]. Recently it also has been shown that spacer material could be obtained in RecBCD independent way with the aid of the Cas4 protein, which is encoded in some type I systems *cas* loci [53].

The second step of CRISPR adaptation is foreign DNA protospacer binding by the Cas1-Cas2 complex. It has been shown that Cas1-Cas2 complex has a strong preference towards double-stranded DNA fragments in vitro [54]. Cas1-Cas2 forms an X-shaped complex with model DNA fragments with the central part of DNA being in a double-stranded form and both ends splayed, forming single-stranded extensions [54, 55]. Binding to Cas1-Cas2 requires that DNA fragment contains a (PAM) [8]. Thus, PAM is essential

both for recognition of protospacer for integration [2] and for target cleavage during interference (see section above) [2, 8]. Yet it remains unclear how Cas1-Cas2 complexes obtain the defined DNA pieces from, presumably, the products of RecBCD degradation. The most recent studies show that when present, Cas4 plays a role in the selection of suitable DNA fragments for the Cas1-Cas2 complex for future integration in CRISPR array [53].

The third step of the adaptation process is an integration of a protospacer into the CRISPR-array. According to the existing model, Cas1-Cas2 cleaves 3'-ends of the protospacer substrate [42] removing two nucleotides of the PAM-complementary sequence leaving exposed a 3'-OH group of a cytosine [54]. Then, Cas1-Cas2 uses these -OH groups to perform two nucleophilic attacks on 5'-ends of the first repeat (leader-proximal repeat) of CRISPR array [54, 56]. Two strands of the previous leader-proximal repeat are separated and form single strands of the first and second repeats of the extended array. The remaining single-strand gaps are then repaired by currently unknown DNA polymerase and DNA ligase [42]. This results in the integration of the new spacer but not the PAM sequence in the CRISPR array and duplication of the repeat.

Other types of CRISPR-Cas systems are less studied yet they possess distinctive features in their adaptation process. For instance, type II systems require their effector Cas9 for a successful acquisition. Presumably, Cas9 is needed to recognize the correct PAM sequence of protospacers selected as the source of future spacers [57]. Newly found and classified type IV CRISPR-Cas system lacks any adaptation genes [5] and it remains unclear what is the mechanism of new spacer acquisition in this case.

1.1.5 Primed adaptation

In type I CRISPR-Cas systems there is also another mechanism of spacer acquisition - primed adaptation. It requires the presence not only of the Cas1-Ca2 complex but also of the Cascade charged with crRNA with a spacer matched or partially matched to foreign DNA [58].. Primed adaptation is much more efficient than naïve adaptation and is highly selective for foreign DNA with new spacers orientation matching that of the priming protospacer recognized by the crRNA.

The detailed mechanism is yet not well understood, and several hypotheses are being considered. The most promising interference-based model proposes the kinetic explanation of primed adaptation [59] and supposes that primed adaptation occurs through a similar mechanism as a naive adaptation except that the spacer source and production mechanism is different. When the Cascade binds to a perfectly matched protospacer rapid degradation of foreign DNA follows [60]. But upon binding to a mismatched protospacer the foreign DNA degradation becomes slower. During normal interference with matched targets, the degradation is so fast that it cannot trigger primed adaptation [61]. Mismatched targets are degraded over an extended period of time and concentration of foreign DNA fragments (generated by Cas3) becomes high enough so they can be picked by the Cas1-Cas2 complex and integrated into the array [62]. Alternative conformational-control model proposes that primed adaptation occurs by a completely different mechanism. It is supposed that binding of the Cascade complex to a spacer with a mismatch could lead to an alternative "open" conformation with the target, compared to “closed” conformation on fully matched target [63]. This open conformation can recruit the Cas1-Cas2 complex which leads to a direct transfer of new protospacers from foreign DNA to Cas1-Cas2 [64].

1.2 CRISPR-Cas systems targeting viruses

1.2.1 Arms race and co-evolution

Since the main function of CRISPR-Cas systems is adaptive immunity and anti-viral defense, the CRISPR-mediated co-evolution of prokaryotes and their viruses was extensively studied both experimentally [65] and computationally [66] over the recent years. Currently, the interaction between viruses and prokaryotic cells is viewed according to a Red Queen hypothesis [67] or evolutionary arms race. Being a fundamental evolutionary concept it proposes that tightly connected species such as predators and their prey or competitor species survive in constant evolution in an endless process of trying to outcompete each other. In the context of CRISPR-Cas system as an anti-viral defense and viruses as a predator of prokaryotes on a short-term scale, this would correspond to a constant acquisition of new spacers by CRISPR-Cas systems and mutation of corresponding protospacer regions in viruses.

While conceptually this arms race between host cells and viruses is straightforward, there is a set of details in the CRISPR-Cas mechanism of action that affects the process, balance, and dynamics of the arms race.

As it was mentioned above viruses can escape targeting by CRISPR-Cas systems through mutations in protospacer regions or their PAMs. It has been shown that two regions are most important in protospacer recognition, which thus should be most vulnerable to mutations in viruses escaping CRISPR-Cas action. These are PAM sequences required for self versus non-self recognition and "seed" region that is important for initial paring of crRNA spacer and target protospacer. It has been shown that single mutations of PAM and most positions of the "seed" result in a drastic increase of effector complex dissociation constant from targets thus abolishing CRISPR-Cas interference. On the other hand, mutations in other regions of spacers lead only to a moderate dissociation constant increase or occasionally have no effect on

spacer-protospacer binding [23]. Yet, strikingly, it was also shown that PAM sequences are not avoided and often are not even underrepresented in viral genomes. Seemingly, the pressure of CRISPR-Cas systems on viruses is not strong enough to promote avoidance of PAMs or, alternatively, PAMs cannot be avoided because of their very short size and also functional redundancy in their sequences [68]

The second complexity in the arms race of viruses and CRISPR-Cas systems is based on the fact that not all spacers are equally beneficial for the host. While any perfectly matched spacer originating from a protospacer with consensus PAM protects the cell from next viral attacks, some spacers are more important than others in the evolutionary perspective. Typically they correspond to more conserved parts of viral genomes that could not be mutated without a significant damage to viral fitness. This is supported by the fact that there are conserved spacers that are maintained in different host subpopulations, strains, or even closely related species. While other factors may contribute to these effects, such as genetic drift, the main source of this universal presence is considered to be selective sweeps [69]. Thus, this functionally splits the CRISPR spacer arrays into two parts: a variable part consisting of unique newly acquired spacers and a conserved part that contains conserved spacers - the effect reported as 'trailer-end conservation' [69, 70, 71, 72, 73]. There are various additional mechanistic details that might come into play. For instance, it has been shown that spacers acquired from the T5 bacteriophage have a strong bias towards the proximal end of its linear genome. The putative reason for that is the location of pre-early viral genes - the genes that are responsible for hijacking the host machinery for virus needs [74]. Thus, targeting these genes gives a significantly higher chance of survival for the host.

CRISPR-Cas systems are somewhat unique in terms of their evolution process. It is commonly accepted that species evolution is Darwinian - the

changes in the progeny occur stochastically and then they are fixed or not in the population under the pressure of environment by survival and reproduction of the fittest individuals [75, 76]. However, with the discovery of CRISPR-Cas systems, it was noticed that their evolution is strictly speaking not Darwinian [77]. Instead, CRISPR-Cas share lots of features with Lamarckian evolution - the concept proposed prior to Darwin's theory and assuming that environment pressure directly affects the changes in the organism phenotype which is then fixed in the progeny. The original point of view of CRISPR-Cas was that it is a unique example of Lamarckian evolution [77]. The current view of the CRISPR-Cas evolutionary process is more complex, and it is considered that it exhibits features of both Lamarckian and Darwinian evolution [77, 78, 79, 80] (see fig. 1.4). Indeed, if we assume that spacers are acquired every time the cell encounters viral particle this type of evolution would be Lamarckian. On the other hand, if spacers are uptaken completely randomly and then cells that have acquired a spacer capable of interference are selected as gaining a distinct advantage - this would count as a Darwinian evolution. As discussed above CRISPR-Cas systems acquire spacers randomly but with a bias towards capturing spacers from viruses. This gives a continuum of evolutionary models with a different degree of "Darwinicity" and "Lamarckicity" based on the extent of bias towards viral DNA [79].

Another paradox related to CRISPR-Cas evolution is the adaptation-survival paradox. When a virus attacks a cell that holds a CRISPR-Cas system there could be two scenarios. The first scenario is the presence of a spacer that targets such a virus. In this case, the cell will destroy it and also could obtain new viral spacer. On the other hand, if the cell did not have a spacer against the virus it could acquire a viral spacer, but typically lytic viruses trigger cell death early in infection, making this outcome highly unlikely. Thus, the paradox occurs - cells seemingly could not inherit spacers obtained by their

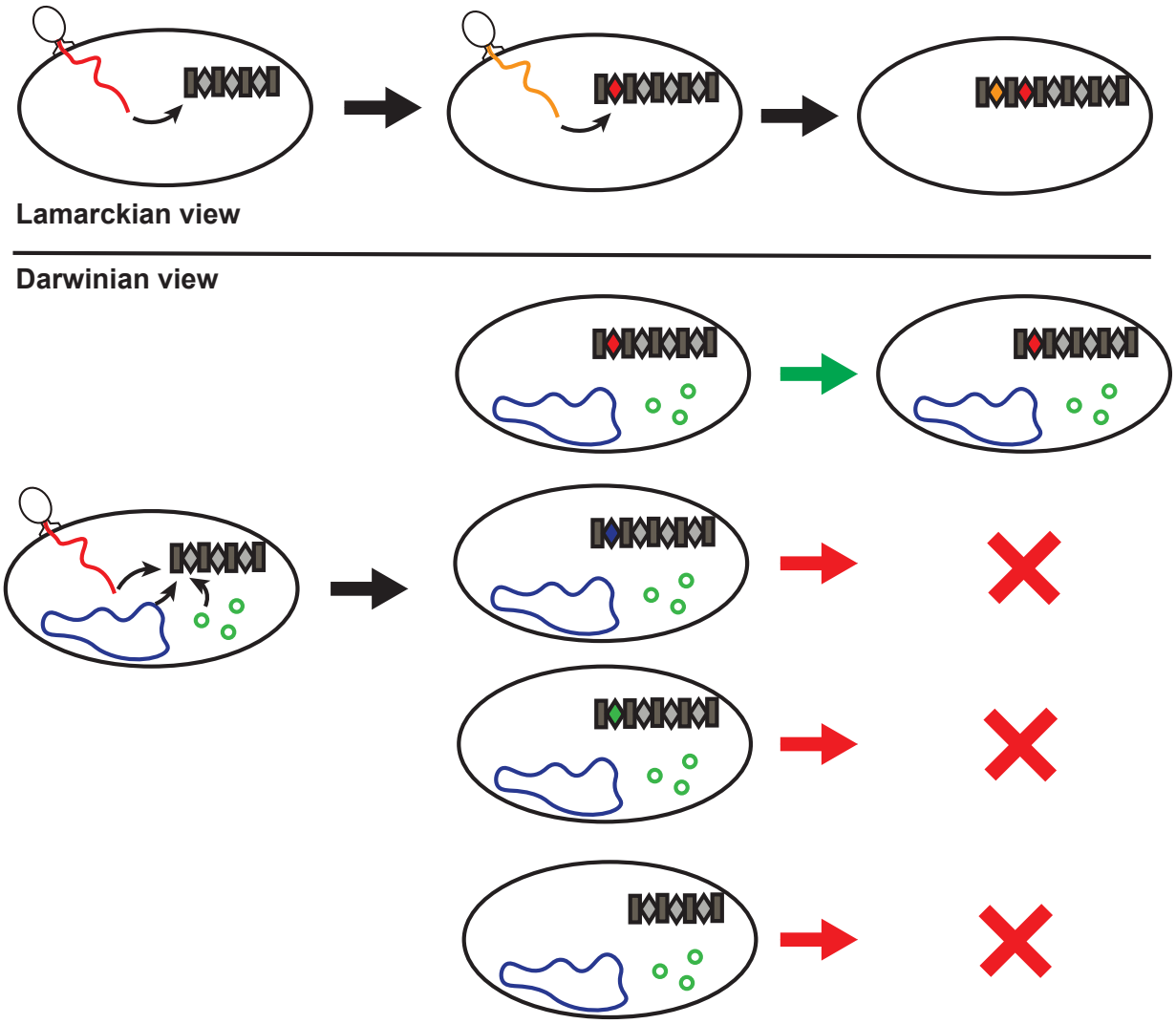

FIGURE 1.4: **Lamarckian and Darwinian evolution of CRISPR-Cas array.** (A) Lamarckian view on CRISPR-Cas evolution. Each viral attack results in spacer acquisition (red and orange) that is being inherited in future generations. (B) Darwinian view on CRISPR-Cas evolution. Random spacer acquisition from different origins (red for viral, blue for host genome, green for plasmids) leads to survival only of those cells that acquired spacers from viruses.

infected ancestors for they are destined to die. For a long time, it was unclear how this paradox is solved and it was proposed that in this case, the survival occur randomly through CRISPR-unrelated mechanisms [79]. Only recently it has been proposed that this paradox could be resolved assuming that spacers could be acquired from defective phages - phages that lost parts of their genome and could not cause cell death. These defective phages could still be a source of spacers and since they share most parts of the genome with functional phages, spacers obtained from these phages shall provide immunity from infection. It has been confirmed experimentally that spacer acquisition from defective phages that are always present as a small fraction in the phage population is highly preferable [81]. However, at the same time, it has been shown that this mechanism does not contribute to all naïve adaptation [81]. Thus, it remains unclear what is the mechanisms that contribute to the rest of the naïve adaptation.

The final level of complexity of the virus-CRISPR-Cas arms race was added with the discovery of anti-CRISPR proteins encoded by phages [82, 83]. They are encoded in specific loci that allow viral infection of cells with perfectly matching spacers. Being a novel class of genes only a few of them have discovered [84, 85]. There are two typical mechanisms of actions [82]. Small acidic anti-CRISPR mimic DNA; they bind CRISPR effector and block it from binding to target DNA [86, 87]. Alternatively, some anti-CRISPR systems function as nuclease inhibitors binding to the Cas3 protein of type I system or to the HNH endonuclease domain of Cas9 [88, 89]. This does not prevent the binding of effectors to the target DNA but inhibits the target DNA degradation. Anti-CRISPR systems lead to the arms race between viruses and host cells leading to rapid evolution and diversification of CRISPR-Cas systems [82].

1.2.2 Altruistic behavior and abortive infection theory

CRISPR-Cas systems are sometimes viewed not as an individual defense system but a population-level defense system and compared with abortive infection systems [90]. Abortive infection systems (Abi systems) were firstly found more than 50 years ago and since then more than 20 different kinds of systems with known or unknown mechanisms have been revealed. Abi systems are prokaryotic anti-virus defense of quite a unique nature [91, 92, 93]. Abi systems do not save the cell from the phage but to the contrary trigger 'programmed cell death' upon the infection. It has been proposed that CRISPR-Cas systems also operate as an abortive infection mechanism. This could work on two different levels. On the first level, CRISPR-Cas could function as proper Abi systems triggering 'programmed cell death' of infected cells when viral DNA is recognized by CRISPR-Cas [74]. On the other hand, CRISPR-Cas systems could act in a more direct way of cutting the invader DNA and lowering the virus progeny quantity, while nor rescuing the infected cell [60]. Such systems are very unique from the evolutionary perspective as their evolution goes beyond traditional Darwinian evolution and involves the evolution of an altruistic behavior. While each individual (in this case a prokaryotic cell) does not survive, such altruistic systems become beneficial for the whole population which is mostly monoclonal. It has been shown that such systems evolve on the balance of the population benefit and individual loss. This scenario contradicts the new gene-based evolution paradigm where the evolution of each individual gene or functionally related gene complex should be viewed separately of the individual evolution and population evolution [94].

1.3 CRISPR-Cas systems targeting plasmids

Alongside viruses as their primary aim, CRISPR-Cas systems can specifically target plasmids. The bias towards plasmid targeting is probably based on the fact that CRISPR-Cas spacers are better acquired from replicating DNA. Since plasmids are typically replicating much more frequently than the host genome it leads to a bias towards plasmid-targeting spacers (and, ultimately, plasmid destruction) compared to self-immunity. It remains unclear whether this feature of CRISPR-Cas systems has some evolutionary advantage or it is an unavoidable cost paid in order to have efficient protection from viruses.

1.3.1 Role of plasmid targeting in nature

Horizontal gene transfer (HGT) is the ability to transfer some genome elements from one organism to another [95, 96]. It is observed mainly in microorganisms and was discovered more than half a century ago [97]. While originally it was thought that HGT plays a minor role in prokaryotic evolution, later it was found that up to 30% of genetic material can be transferred through HGT and HGT obviously played a critical role in prokaryotic evolution [98]. Nowadays it is argued that even the concept of "tree of life" as a series of species bifurcation should be abandoned towards a "web of genes" when each gene evolution should be viewed independently [94, 99]. Three main mechanisms of HTG are conjugation, transformation, and transduction [95]. Conjugation occurs when cells contact each other, forming junctions allowing the genetic material transfer. It is typically mediated by conjugative plasmids that could both carry the conjugative systems and be transferred through the junction [96, 100]. Transformation is an uptake of genetic material that is floating freely in the environment [101]. The main source of such material is lysed cells [102]. Transduction is a transfer of genetic material by

viruses that results in integration of viral genome or non-infectious DNA carried in viral particles in the host genome [95]. Thus, during HGT host cells mainly rely on plasmids and viral agents as the genomic material carriers and plasmids play a crucial role in HGT.

It has been shown in numerous experiments that CRISPR-Cas systems that target plasmids prevent HGT [103, 104] both through conjugation [103] and natural transformation [105, 106, 107]. Moreover, the only way for the plasmid to escape the elimination by CRISPR-Cas system targeting that plasmid is through mutation of protospacer, spacer or other components of CRISPR-Cas machinery (see fig. 1.5) [106]. While an acquisition of new spacers from plasmids is less studied, a bioinformatic study has shown that a large fraction of spacers originate from plasmids [108], suggesting that CRISPR-Cas is limiting HGT.

On the other hand, despite strong evidence of efficient inhibition of HGT by CRISPR-Cas on a timescale of individual cells and cell populations, there is a lack of evidence of CRISPR-Cas efficiency against HGT on evolutionary timescales. Bioinformatic analysis of the correlation between spacer acquisitions and HGT during evolutionary history revealed a lack of correlation between the activity of CRISPR-Cas and quantity of genes acquired through HGT. While there are several putative explanations, including inefficiency of CRISPR-Cas against very high exposure to mobile genetic elements, there is yet no good explanation of this observation [109].

1.3.2 CRISPR-Cas system spacer diversity and spacer origin

There is an observed intriguing diversity in the spacer space. There are multiple bioinformatics studies that have analyzed the spacer space of various species [73, 110, 111]. It was revealed that typically different sprains of

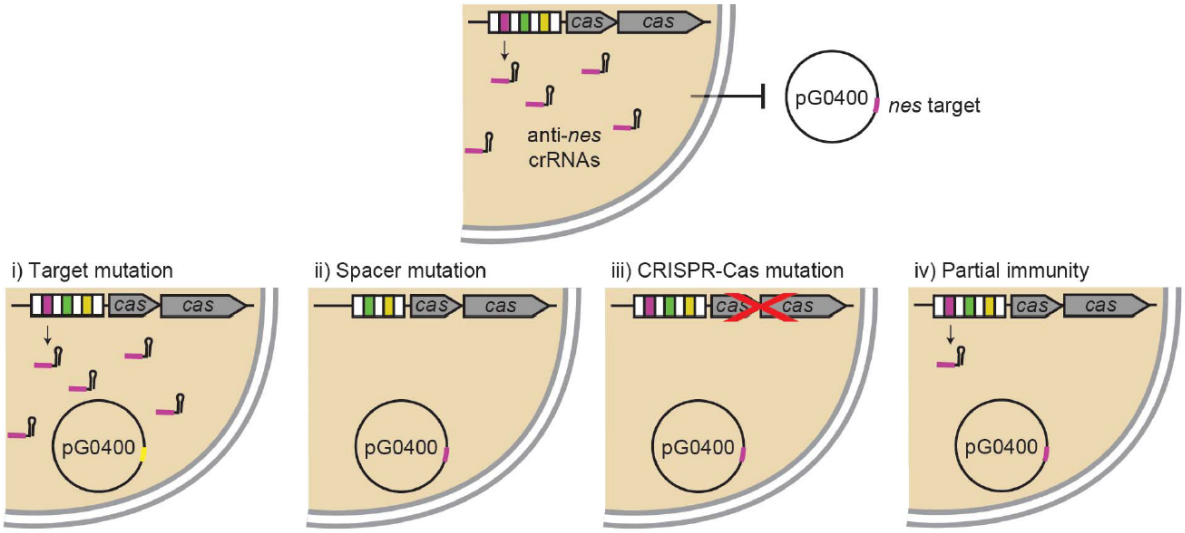

FIGURE 1.5: **Result of the interaction of plasmid and CRISPR-Cas system in the "offer they can't refuse" experiments.** The plasmids survived under the pressure of CRISPR-Cas systems targeting the plasmids through the mechanism that falls in one of four cases. i) Plasmid target protospacer mutation that leads to loss of recognition and the following interference. ii) Alternatively, spacer in the CRISPR array was mutated or lost leading to loss of immunity. iii) One of Cas proteins in the host cell mutated leading to a dysfunctional CRISPR-Cas system. iv) There was a mutation in the CRISPR-Cas regulation leading to partial immunity. The figure was reused without changes from [106] at PloS Genetics under Creative Commons Attribution 4.0 International.

the same host species have very diverse spacer composition. The main intrigue is relater to the origin of CRISPR spacers. While the main function of CRISPR-Cas systems is anti-viral defense it should be expected that all the CRISPR spacers should be originated from viral genomes. However, a series of bioinformatic experiments surprisingly reveals that it is very far from truth [112, 113, 114]. Several studies revealed the variety of spacer origin: alongside with virus sources some spacers target plasmids, genomes of other species, unidentified pangenome material or even self-targeting [93]. However, from 70% to 95% spacers still remain unidentified origin and were not mapped on any known sequence. It remains unclear what is the source of this so-called "dark matter" of spacers. There are several hypotheses: those spacers correspond to rapidly mutated viral protospacers which could not be identified; those spacers correspond to undiscovered viral species; those spacers are generated through some unknown mechanism or correspond to some other functions.

1.4 Other functions of CRISPR-Cas

Analysis of different aspects of the CRISPR-Cas system has revealed that in addition to their function as the prokaryotic immune system there likely are secondary functions. Given their nuclease activity, they putatively hold a genome editing and transcription regulation functions. These unique features could also be utilized in various model systems and genome editing tools [93, 115, 116].

Most commonly reported secondary function of CRISPR-Cas systems is the regulation of gene expression. There are two putative mechanisms that have been proposed which are probably complementary and may function

simultaneously. The first mechanism is regulation based on partial complementary between crRNA and cellular target. While the details of this regulation remain unclear, it is proposed that this mechanism is similar to the mechanism by which microRNA (miRNA) regulate gene expression in eukaryotes [115, 117]. While perfect match to a host target locus has a high probability to result in self-immunity and degradation of self-DNA, CRISPR arrays rarely hold spacers identical to host genes. However, partial matches are often observed. It is proposed that effector complexes charged with these partially matching crRNAs could bind to mRNA resulting in inhibition of translation by the physical arrest of a ribosome, similarly to miRNA mechanism of gene regulation [117]. It is also possible that CRISPR-Cas effectors arrest transcription by binding to a partially matched gene without promoting its cleavage. This mechanism was shown as a putative regulatory mechanism of biofilm formation in *Pseudomonas aeruginosa* [118].

The second mechanism is targeting mRNA and its direct degradation. While most CRISPR-Cas systems target DNA, some, such as type III [35, 119] or type VI [10, 41] target RNA. In the case of such systems, a perfect match of a spacer to a gene part will result in a knockdown of the gene by mRNA degradation without causing autoimmunity.

CRISPR-Cas might also play a role in DNA repair mechanisms since Cas1-Cas2 proteins are expected to interact with the RecBCD system [43], which is involved in DNA repair [44]. It has been shown that Cas1-knockout strains showed higher sensitivity to DNA-damaging agents. Cas1 was showed to play a role in the resolution of Holliday junctions that occur during DNA recombination [120].

Last but not least is the possibility of CRISPR-Cas induced genome alterations by direct self-targeting. This occurs if a CRISPR spacer is targeting some regions of the host genome that could not be distinguished by the existing self- versus non-self differentiation mechanisms. While in most cases

such targeting results in high toxicity and inevitable cell death, it can also trigger genome rearrangements that could lead to rapid evolution [121].

1.5 CRISPR-Cas costs

Currently, it is generally agreed that there are downsides of holding a CRISPR-Cas system (sometimes referred to as CRISPR-Cas fitness cost), though there is no agreement what is the main source of such cost. There are several ideas being discussed. They include genomic burden, the cost of maintenance of the *cas* genes, potential auto-immunity, and blockage of beneficial HGT [122, 123].

The genomic burden is due to the maintenance of extra genetic material in the form of CRISPR-Cas loci. The prokaryotic genome has a high bias for deletions for the sake of economy and efficiency [124]. However, this should not have a large effect on CRISPR array evolution as the size of even largest CRISPR systems is less than 1% of the genome [125].

Targeting self DNA by spacers is causing self-immunity and is very toxic to the bacteria [121], and such spacers are clearly avoided. Self-immunity cost is a relatively complex issue for CRISPR systems as a mechanism of avoiding self-immunity is not well studied and seemingly differs in different CRISPR types [79]. It has been shown that Type II-A CRISPR system acquires spacers both from foreign DNA and self DNA [126]. At the same time in system I-E, there is a preference towards foreign DNA due to acquisition preference towards rapidly replicating genetic elements [6, 43]. Alternatively, CRISPR-Cas expression could be repressed in the normal living cells and could be induced only in the course of infection [127]. These mechanisms are not mutually exclusive and have the potential to jointly reduce the self-immunity cost. Be as it may, this cost seems to depend on the activity and mechanism of the initial stage of spacer acquisition and does not depend on the length of

the spacer array. If a single spacer is acquired from self-genome it will result in almost inevitable cell death [121].

Experiments have shown that CRISPR-Cas systems inhibit plasmid transformation and have a potential to limit horizontal gene transfer [103] which plays a crucial role in prokaryotic evolution. It has been shown that the introduction of a strongly beneficial plasmid that is targeted by CRISPR leads to loss of CRISPR [106] (fig. 1.5). Indeed it has been shown that the marks of HGT attempts could be seen in CRISPR of the archaeal genomes [108]. However, a large-scale bioinformatics study has shown the increase in the length of the array is not correlated with the decrease of new genes acquisition [109]. Thus, this cost (if it exists) is associated with the activity of the CRISPR system, and not with the length of the array.

Another simple yet important cost is the cost of maintenance of the CRISPR-Cas system i.e. crRNA synthesis, effector synthesis, the energetic costs of new spacer acquisition etc. A recent study has shown [123] that this cost indeed exists, and is mainly related to the production of Cas proteins and is not affected by the number of spacers in the CRISPR array.

1.6 Modeling approaches to study CRISPR-Cas systems

Since this work is focused on modeling of CRISPR-Cas systems, we need to discuss the approaches that have been used to do so in the past. There are multiple computational works that explored CRISPR-Cas phenomenon in order to unwind the co-evolutional complexity of the system [66]. There are two main approaches or techniques that are commonly used in the literature. The first technique is modeling using a system of differential equations (ODE-models) where each population or subpopulation of host cells

and viruses is described by its own differential equation. In this framework, the dynamic of each population is deterministic and continuous. The second approach is agent-based stochastic models or individual-based models (ABS-models) that treat each individual cell or virus particle as an independent object.[128]. Some works used a specified two-dimensional spatial stochastic models in order to capture the spatial features of the process [78, 129]. Some works incorporated the best features of both ODE and ABS approaches in order to capture the randomness of the underlying processes of spacer acquisitions and protospacer mutations yet keep the model computationally efficient [80, 130, 131, 132].

The first mathematical model that incorporated the features of CRISPR-Cas was developed in 2010 by Levin [133] as a simple ODE Lotka-Volterra-like system that incorporated binary CRISPR-based immunity: The host cells were either non-immune or fully immune to the viruses. The model was developed in order to assess the potential efficiency of CRISPR-mediated phage resistance and compare it with other types of resistance, for instance, envelope resistance. In other words, it addressed the question when the CRISPR-Cas system should be the dominant host defense system and why some procaryote species has CRISPR-Cas while other does not. It has been shown that there is some parameter space where holding CRISPR-Cas gave an evolutionary advantage for the host cells, however, this parameter space became narrower if an alternative immunity mechanism existed. Further analysis by the same group used both computational and experimental techniques in order to validate predictions and extend the model [106, 134]. In particular, they were focused on the host-virus co-evolution and extinction. These works show that a single spacer acquisition is not enough and in order to be fully immune to viruses host cells should hold multiple spacers (two in case of *Streptococcus thermophilus* DGCC7710) that should be obtained through first- and second-order spacer acquisition [134]. At the same time, if there is some

subpopulation of the host cells that remain vulnerable to the viruses, the host cell population cannot eliminate viruses, leading to their coexistence[134]. They have shown that the presence of CRISPR-Cas in the system promotes arms race and coexistence of virus and host. Further work analyzed the fitness cost of CRISPR-Cas and the ways the host cell deals with CRISPR-Cas targeting beneficial agents. They tested the "offer they can't refuse" type of system - introducing a plasmid that provides antibiotic resistance that is targeted by CRISPR-Cas [106]. They have shown that such a system operates in a tight balance of functional and knocked-out CRISPR-Cas depending on the host cells mortality rate due to viral infections and other causes [106]. A further analysis on a similar topic was performed by Weinberger et al., counting costs and benefits of CRISPR-Cas system [135] (fig. 1.6). They used an ABS-model of virus-host coevolution including CRISPR-Cas system costs and viral mutation rate as the system parameters. They revealed that there is a sharp edge separating the host population with inefficient CRISPR-Cas from the population with efficient CRISPR-Cas. It was coupled with bioinformatic analysis showing that other environmental factors such as temperature can affect this balance [135]. Further analysis linked virus mutation rate and viral population size, providing the virus population diversity threshold that is limiting the efficiency of CRISPR-Cas systems [136]. At the same time, they provided an alternative more simple explanation of temperature effect on CRISPR-Cas systems suggesting that the direct factor that increases the efficiency of CRISPR-Cas systems under such conditions is the low population size which itself is affected by extremal temperature conditions. In the work of Berezovskaya et al. they took a different approach, applying a bifurcation analysis to an ODE model of host-virus co-evolution [137] to further explore potential regimes of such co-evolution. This model has shown that alongside with trivial equilibrium (i.e. elimination of either host or virus), the system may also exhibit non-trivial stable equilibrium or oscillations. However, it

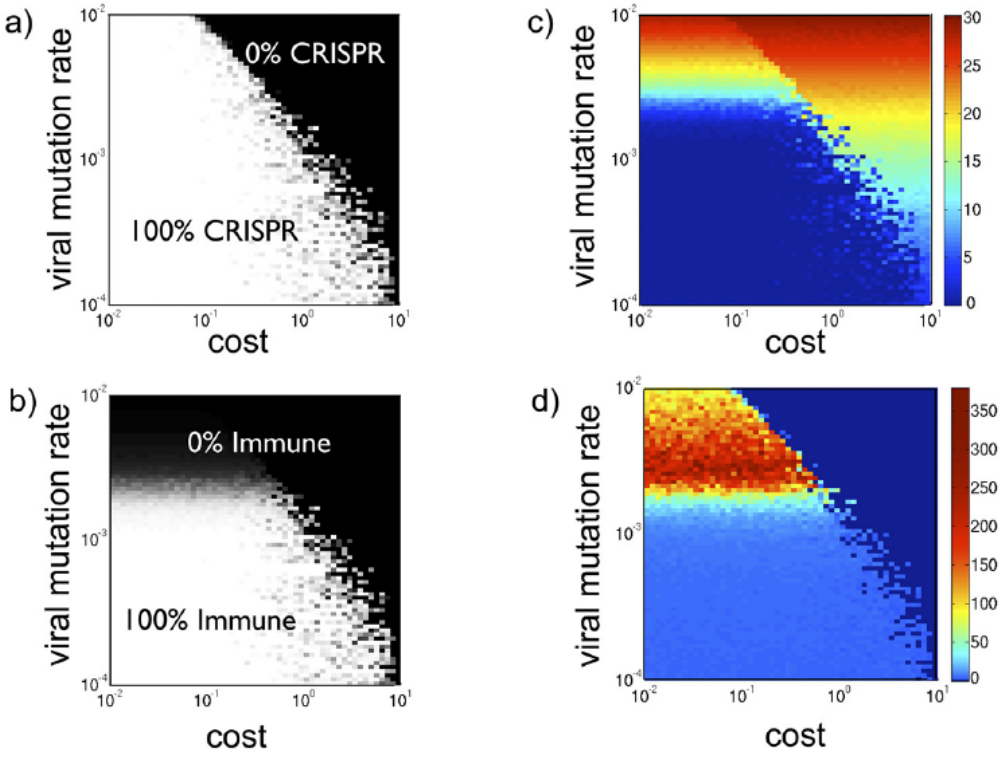

FIGURE 1.6: **Cost-benefit analysis of CRISPR-Cas prevalence** (A) Threshold of CRISPR-Cas system presence based on its cost (B) Efficiency of CRISPR-Cas system (C) Diversity of the viral population when suppressed by CRISPR-Cas system of the hosts (D) number of spacers in the CRIPSR array. The figure was reused without changes from [135] in mBIO journal under Creative Commons Attribution-Noncommercial-ShareAlike 3.0 Unported.

has been shown that oscillations are often quasi-chaotic, making impossible to make any valid prediction on the result of virus-host co-evolution.

Another direction of modeling is focused on various aspects of diversity maintenance, co-evolution, and co-existence of virus and host sub-population. While natural virus-host populations seem to co-evolve [66, 73, 138], it remained unclear how this co-evolution occurs in case of Lamarckian-like setting of CRISPR-Cas. In the works of Childs et al. they explored such systems in a model that combine ABS and ODE model features [80, 130].In their work, they explored the driving forces of the coexistence of virus and host

and how this coexistence affects the composition and dynamics of spacer arrays. They have shown that such co-existence and long-term maintenance of multiple subpopulations of hosts and viruses occur naturally through incomplete sweeps by single strains and temporary advantage of the older strains [80]. Further analysis was focused on the hedge immunity as they introduce the concept of distributed immunity and Population-wide Distributed Immunity (PDI) as a measure of overall immunity of all host subpopulations in [130]. They proposed that the host population is presented by genetically diverse but phenotypically similar subpopulations that function in a coalition against viral attacks [130]. Tightly related to this works is an article by Weinberger et al. who assessed the trailer-end conservation phenomenon (see Arms race and co-evolution section) both bioinformatically and through modeling [69]. They have shown that unidirectional spacer acquisition and selective sweeps naturally lead to loss of spacer diversity of trailer-end. At the same time, another model has shown that only a few leader-end spacers contribute to immunity while trailer-end spacers are almost useless [80].

In other models Haerter et al. also studied co-existence, co-evolution, and diversity of viruses and host cells but in a setting driven by spatial effects [78, 129]. They have shown that spatial distribution of host cells and virus population leads to a co-existence of host cells in the average viral species number that exceeds the capacity of CRISPR-Cas array due to a non-universal distribution of viruses [129]. Also, the viral infection leads to a spatial distribution of cellular subpopulation with different array composition. The cells on the edge of the infection acquire new spacers from viruses which becomes dominated by other spacers as cells move "inside" the population [129]. The further work analyzed whether the CRISPR-Cas evolution is Darwinian or Lamarckian. Surprisingly, it has revealed that spatial distribution of viruses makes the overall CRISPR-Cas spacer evolution more Lamarckian and at the same time more efficient [78].

Overall the evolutionary models of CRISPR-Cas systems have shown a somewhat limited applicability: While being able to prove some fundamental concepts, certain models fail to capture the complexity of the whole CRISPR-Cas system. Some modeling conclusions contradict each other as a result of different design and complexity depth of the models. At the same time, the inner stochasticity underlying most of the processes of spacer acquisition, protospacer mutation etc. leads to the whole process lacking deterministic predictive power [66, 137].

Chapter 2

Introduction

CRISPR-Cas systems are a relatively novel discovery proven to be extremely useful as a genetic engineering tool and attracting constant attention from the modern scientific community. Being extensively studied in the last decade, CRISPR-Cas phenomena still hold numerous paradoxes and undiscovered mechanisms. Unique features of CRISPR-Cas such as semi-Lamarckian evolution makes it an extremely interesting topic to study, in particular through computational methods.

Computational modeling of CRISPR-Cas generally falls into two categories: simulation of the host-phage co-evolution and assessment of CRISPR-Cas efficiency as a whole. Co-evolutionary types of works asked questions like: Can the host and viruses coexist in the presence of CRISPR-Cas system in the host and what are the conditions that lead to coexistence or extinction of one player [80, 130]? How spatial distribution of host cells affect the co-evolution and spacer acquisition [78, 129]? What drives the spacer selection [69]? At the same time, CRISPR-Cas efficiency and cost-benefit analysis typically addressed the questions of the prevalence of CRISPR-Cas system and the conditions when the CRISPR-Cas system is beneficial or not [69, 106, 133, 134]. The array composition and number of spacers is also sometimes assessed in these types of models [69], but it is also performed using co-evolution approach. While proving several fundamental concepts,

these works yet remain disjointed and showing different, often contradictory results. At the same time, most of the present models use the similar methodology, being co-evolutionary ODE or agent-based models with various features of CRISPR-Cas system being included. Researchers working on modeling CRISPR-Cas systems still look for new approaches and techniques in order to improve our understanding of their functions and mechanisms

In our work, we focused on the application of computational techniques that were rarely used in the works of other research groups on the modeling of CRISPR-Cas systems in order to explore the field from the different perspective. We believe that straightforward co-evolution models of such a complex system as CRISPR-Cas have limited applicability and do not allow one to make a quantitative and qualitative assessment of CRISPR-Cas system behavior.

In this work, we computationally assessed several aspects of CRISPR-system and computationally analyzed how its characteristics affect array composition, cell-virus interactions, and cell-plasmid interactions. We specifically focused on spacer array-independent characteristics such as efficiency of CRISPR interference and number of CRISPR effector complexes as they are poorly studied computationally yet crucially affect all array-dependent processes. In particular, we focused on two related topics: we estimated the optimal number of spacers that maximizes the cellular survival and investigated the dynamics of the CRISPR-Cas interaction with foreign genetic elements.

To assess the optimal number of spacers in the array, we analyzed how the survival of a host cell under a series of viral attacks depends on the array composition. Instead of the evolutionary dynamics of CRISPR-Cas and viruses, we focused on the array composition that optimizes host survival at any given time. This, turn around the question of virus-host co-evolution, focusing on the goal of the evolutionary process rather than the process. We

also have incorporated several unique features such as unequal crRNA production from the spacers in different positions that was proven to play a crucial role in CRISPR-Cas immunity yet poorly studied. While rather idealized, this model provides a different viewpoint on the whole process of CRISPR-Cas immunity and on the result of other models.

Further, we focused our analysis of the details of the interaction of CRISPR-Cas and plasmids. While other works typically consider the interaction of CRISPR-Cas and foreign genetic elements at the population level omitting most of the kinetics details, we particularly focused on the molecular kinetics of plasmid replication and CRISPR-plasmids interference events. We developed a stochastic model of the plasmid replication, CRISPR interference, and cell division, that assesses the plasmid distribution in the host cell population. Followed by a set of experiments, we managed to explain and prove that interactions between CRISPR-Cas and foreign genetic element are far more complex than it was thought before and lead to non-deterministic outcomes.

Chapter 3

Optimal number of spacers in CRISPR array

3.1 Introduction

CRISPR-Cas systems provide prokaryotes with adaptive immunity against viruses and plasmids by targeting foreign nucleic acids [1, 8, 139]. 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, 140]. 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 [111, 113, 114]. The ubiquity of CRISPR-Cas systems in archaea compared to less than 50% presence in bacteria is also not well-explained [140, 141]. Evolutionary reasons for a plethora of distinct CRISPR-Cas systems types, often coexisting in the same genome, remain largely unexplored [4, 125, 142]. 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 [70, 125, 142, 143, 144, 145]. 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 [134]. However, even the largest of the CRISPR systems contribute only 1% to the total size of a prokaryotic genome [125], 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 [103, 146], none of them significantly depends on the number of the CRISPR spacers [109, 123, 125].

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, [80], 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 [136, 147]). 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 a 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 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. According to the observed results, the main drivers of the diversity in the optimal number of spacers are the mutation rate of the viral population and the expression level of the CRISPR-Cas system.

3.2 The Model

3.2.1 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. 3.1). 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

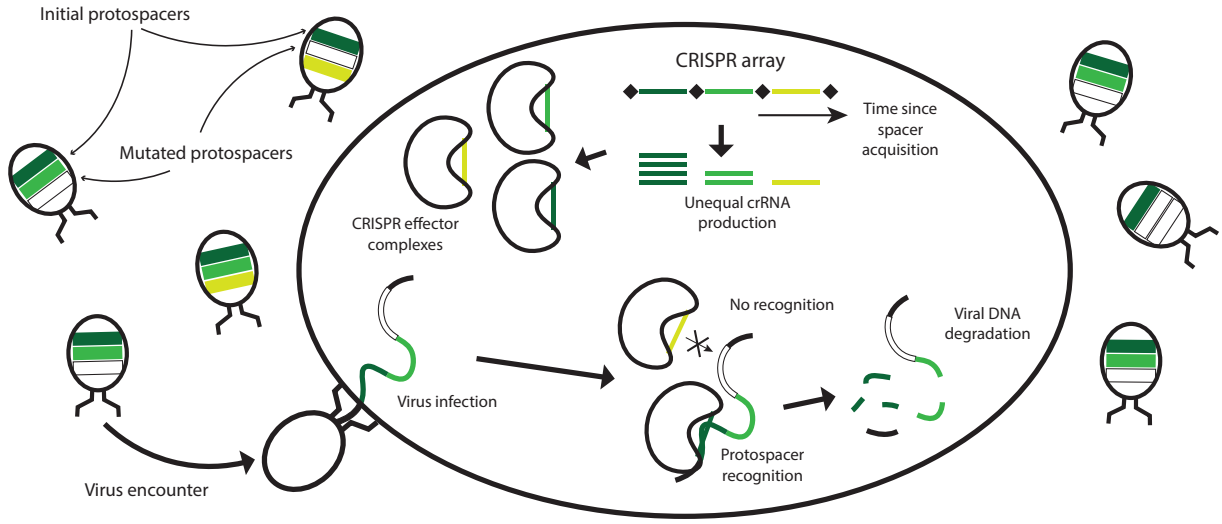

FIGURE 3.1: **Functioning of CRISPR-Cas system in relation to viral attacks.** Three spacers are colored according to their age from the time of their acquisition, from dark green marking the youngest (the most recently acquired) spacer to yellow marking the oldest one (which was acquired the earliest). 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 the destruction of viral DNA while the latter leaves it intact.

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 systems to recognize viral DNA [134]. 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 [148, 149]. As a result, spacers in natural arrays are ordered according to their age, with more recently acquired spacers located closer to the 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 [150]. 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 [1]. While the reality is more complex and certain mutations in protospacers do not preclude recognition by the appropriately

charged effector [23], mutations in protospacer adjacent motif [151, 152] or seed region [23] 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 [80, 136]. 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 the 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. 3.1 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 [23, 151, 152].

- Probability for a CRISPR effector complex to contain crRNA with a particular spacer decreases exponentially with the age of the spacer. The higher abundance of leader-end, thus younger, spacer crRNA is observed and the abundance of subsequent spacer crRNA gradually drop [150] we thought this would be a potentially accurate assumption given the potential mechanisms that could contribute to such effect (see section 3.4.3).

- The total number of CRISPR effector complexes in the cell is constant on the timescale of virus attack. While there exists evidence for *cas* genes expression being regulated in vivo depending on the external

conditions [37, 38] and, in particular, triggered by the viral invasion [39, 40], there are no signs that in the actual course of a virus attack or under a typical experimental condition, i.e. in "fully active" state, the levels of Cas proteins change. Thus, the assumption that at given conditions expression levels are constant seems to be reasonable.

- 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 [109, 123].

- 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 stay the same. For the multiple-virus case (see sections 3.3.4, 3.3.5) 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 [61] 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.

3.2.2 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,

$$\frac{dB_i}{dt} = k^+(1 - B_i)C_i - k^- B_i. \tag{3.1}$$

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

$$B_i = \frac{k^+ C_i}{k^+ C_i + k^-} = \left[1 + k^- / (k^+ C_i)\right]^{-1}. \tag{3.2}$$

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 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 [150]. Hence the number of effector complexes C_i with crRNA with spacer i is

$$C_i = C_1 \delta^{i-1}. \tag{3.3}$$

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

$$C_i = C\delta^{i-1}\frac{1-\delta}{1-\delta^S} \tag{3.4}$$

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

Substituting (4.7) into (3.2) produces a complete expression for the binding probability between the ith spacer-protospacer pair,

$$B_i = \left(1 + \frac{1}{\beta}\frac{1}{\delta^{i-1}}\frac{1-\delta^S}{1-\delta}\right)^{-1}. \tag{3.5}$$

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,

$$B = \frac{1}{1 + 1/\beta}. \tag{3.6}$$

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 corresponding 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 corresponding spacer-protospacer pairs, and the probability of survival of viral DNA $P(t)$ decays with a simple exponential kinetics,

$$\frac{dP(t)}{dt} = -aP(t)\sum_i B_i; \;\; P(t) = \exp\left(-at\sum_i B_i\right). \tag{3.7}$$

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

$$I = 1 - P(\tau), \tag{3.8}$$

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 cannot prevent the runaway viral DNA replication

and productive infection. Introducing a dimensionless parameter $\chi \equiv \tau a$, which characterizes the interference efficiency, turns Eqs. (3.8 and 3.5) into

$$I = 1 - \exp\left[-\chi \sum_i B_i\right] = \tag{3.9}$$
$$1 - \exp\left[-\chi \sum_i \left(1 + \frac{1}{\beta}\frac{1}{\delta^{i-1}}\frac{1-\delta^S}{1-\delta}\right)^{-1}\right].$$

3.2.3 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,

$$A_k(t) = \frac{(rNt)^k}{k!}\exp(-rNt), \tag{3.10}$$

where r is a proportionality coefficient considered to be the same for all viruses and N is the 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

$$\sum_{k=0}^{\infty} A_k(t) I^k = \exp[-rNt(1-I)]. \tag{3.11}$$

Here I, defined in Eq. (3.9), 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,

$$E(t) = \prod_j \exp[-rN_j t(1 - I_j)]. \tag{3.12}$$

This is sketched in Fig. 3.2 with the modeling parameters as in table 4.1. The probability of CRISPR interference with a single infection I_j is defined as in (3.9) 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.

3.2.4 Calculation of interference efficiency from experimental data

We use the data from [74], which quantitatively assesses the efficiency of interference of a single-spacer CRISPR system against the T7 phage with a perfectly matching protospacer. The DNA abundance from the protospacer region, which is cut by CRISPR effectors, and the reference unaffected by CRISPR are compared to each other. Since the probability for the viral DNA to survive a duplication cycle is $1 - I$ (see eqs. (3.7) and (3.8)), the number of copies of the protospacer region of viral DNA V_{CRISPR} after ν rounds of duplication is

$$V_{CRISPR} = [2 * (1 - I)]^{\nu}. \tag{3.13}$$

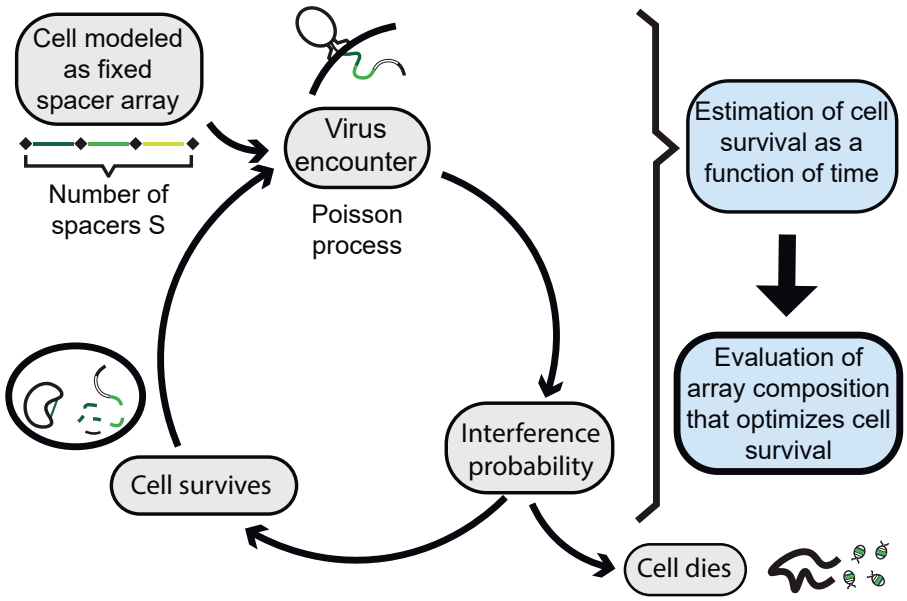

FIGURE 3.2: **Scheme of calculations of the optimal number of spacers in CRISPR-Cas array.** A cell with $S = 3$ CRISPR spacers encounters viruses as a Poisson process with an average rate of rN. During each encounter, there is either a successful interference with probability I or the cell dies with 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.

TABLE 3.1: List of parameters used in the model of the optimal number of spacers in CRISPR array

Model parameter	**Parameter name**	**Parameter description**
S	Number of spacers	Represents the number of spacers in the array of the given cell.
δ	crRNA decay coefficient	Shows the ratio between levels of crRNA originated from i-th and $i+1$-th spacer.
β	Binding efficiency	Represents how well the CRISPR-Cas effector binds to the target DNA, $\beta = 1$ correspond to the occupation of protospacer by CRISPR-Cas effector with 50% of the time.
χ	Interference efficiency	Corresponds to the chance of degradation of target protospacer over the effective time of interference.
rNt	Average viral encounter time	Corresponds to the average time over which the cell will encounter a given number of viruses, taking into account the viral load N and the encounter rate r.
$1-\mu$	Mutation probability	The probability of the protospacer to mutate and become unforgettable over the time of a spacer acquisition.

The CRISPR-free viral burst size and, presumably, the number of copies of reference regions of phage DNA is $V_b \approx 100$ viruses, thus the average number of virus duplications ν is given by

$$2^{\nu} = V_b = 100,\ \nu = \frac{\ln 100}{\ln 2} \approx 6.65. \tag{3.14}$$

The ratio between the amount of DNA from the reference and protospacer regions was reported in [74] to be approximately 100,

$$\frac{V_b}{V_{CRISPR}} \approx 100. \tag{3.15}$$

Thus

$$[2 * (1 - I)]^r \approx 1,\ I \approx 0.5. \tag{3.16}$$

The relation between β and χ that reproduces the interference probability of the single-spacer array from [74] is obtained by inverting the eq. (3.9) and limiting the sum to the first term,

$$\chi = -\ln(1 - I)(1 + 1/\beta) = (1 + 1/\beta)\ln(2). \tag{3.17}$$

3.3 Results

3.3.1 Application: Single viral species

To illustrate and further develop the general statement (3.12), 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 the viral population. In this model, the mutation status of the spacer will be defined as the fraction of mutated protospacers

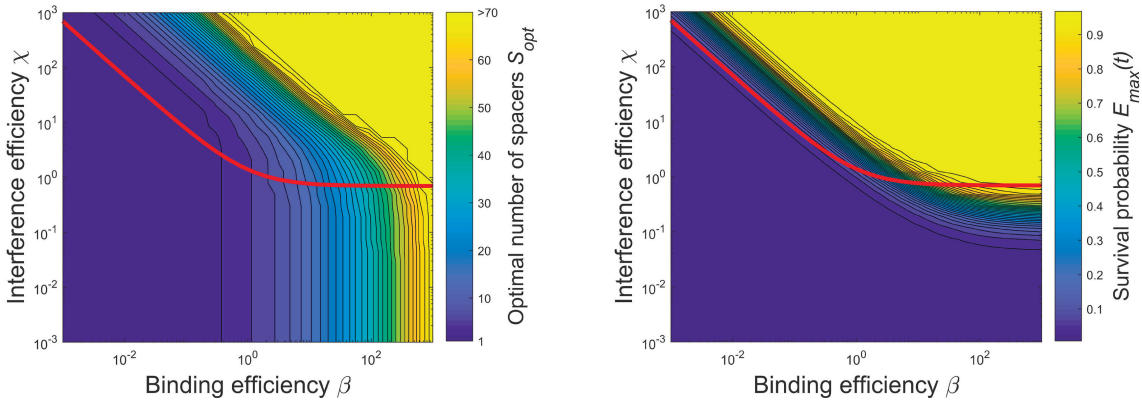

FIGURE 3.3: **The effects of binding efficiency β and interference efficiency χ on CRISPR performance** The optimal number of spacers (left panel) and the corresponding survival probability (right panel) are shown for various β and χ. The probability for the protospacer to remain mutation-free is $\mu = 0.9$ in both plots. Red line corresponds to the values of β and χ satisfying the restriction $I = 0.5$ given by Eq. 3.17.

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. (3.9 and 3.12) and our assumption that a mutation in protospacer renders the corresponding spacer completely inefficient, it follows that the survival probability in such case is

$$E(t) = \exp\left(-rNt\left\{m_1 m_2 \exp[-\chi(B_1+B_2)] + \right.\right. \tag{3.18}$$

$$\left.\left. + m_1(1-m_2)\exp[-\chi B_1] + m_2(1-m_1)\exp[-\chi B_2] - (1-m_1)(1-m_2)\right\}\right).$$

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

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{2}\{1 - m_i[1 - \exp(-\chi B_i)]\}\right)\right]. \tag{3.19}$$

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 (3.19) is the probability for the Poisson process of "failures" of the CRISPR-Cas 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 (3.19) to cases of arrays containing more than 2 spacers, replacing the upper limit of the product by an actual number of CRISPR spacers S,

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{S}\{1 - m_i[1 - \exp(-\chi B_i)]\}\right)\right]. \tag{3.20}$$

The equations (3.12) and (3.20) are universal and are applicable to a variety of scenarios involving CRISPR immunity. For example, (3.12,3.20) can serve as a base for evolutionary dynamics models, where the mutation status of protospacers and the composition of CRISPR array is 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 [74] or the co-evolution of CRISPR individual immunity and altruistic abortive infection mechanisms [115]. However, it is hard to visualize the conclusions that follow from (3.12,3.20) in their general form due to the a large number of generally unknown parameters m_i.

To reduce the number of independent parameters in Eq. (3.20) 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 the 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. While the actual distribution of the time interval between spacer acquisitions holds stochastic nature the older spacers acquisition times follows the normal distribution with the mean of $\mu^{i-1/2}$. This we consider it to be a fair assumption. On the other hand, the factor that may conflict this assumption is the selection of one spacer over another and non-uniformity in the evolutionary distances between different spacer acquisitions. This could

be viewed and analyzed within the framework of this model as several different groups of spacers, for instance, trailer-end spacers that were selected and newly acquired spacers that did not undergo the selection.

Eq. (3.21),

$$E(t) = \exp\left[-rNt\left(\prod_{i=1}^{S}\left\{1-\mu^{i-1/2}[1-\exp(-\chi B_i)]\right\}\right)\right], \tag{3.21}$$

together with the binding probability (3.5), 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. (3.21) the position of the maximum of $E(t)$ is determined by the maximum of the product in the exponent and is independent of rNt.

3.3.2 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. 3.4. For this example, we inferred the interference probability $I_1 \approx 0.5$ of a single spacer array from the experimental data [74] (see S1 Appendix for details). While the exact values of binding efficiency and interference efficiency 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. It is shown in [13] 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 S1 Appendix, section 2 for an example which uses a different pair of β and χ for the same I. The probability for a protospacer not to mutate

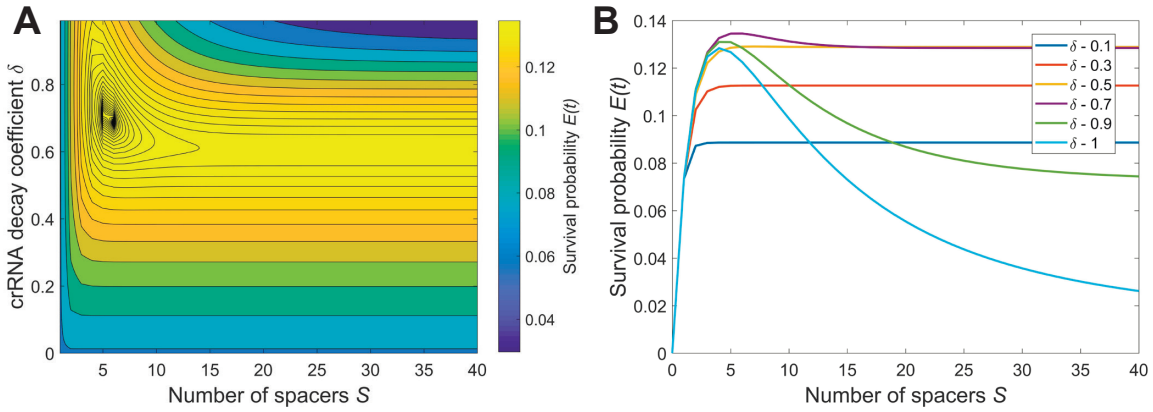

FIGURE 3.4: **Typical survival probability profile.** (A) Plot of survival probability $E(t)$ vs. the crRNA decay coefficient δ and the number of spacers in CRISPR array S. Other parameters are: $\beta = 1$, $\chi = 1.4$, $\mu = 0.9$, and $rNt = 5$. (B) Six curves of $E(t)$ vs. S for various values of δ and same β, χ, m, and rNt as in the panel A.

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. 3.4 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 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 crRNA with older spacers in effector complexes: no matter how long the array is, the 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. 3.5 we show how the plot of the "typical case" shown above in Fig. 3.4 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. 3.5, 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. 3.5, 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. 3.6. As discussed above, higher viral mutation rates lead to lower survival probability and fewer spacers (Fig. 3.6A). 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 I_{max} that at least one of viral protospacers has not mutated.

$$I_{max} = 1 - \prod_{i=1}^{S} (1-\mu)^{i-1/2} \tag{3.22}$$

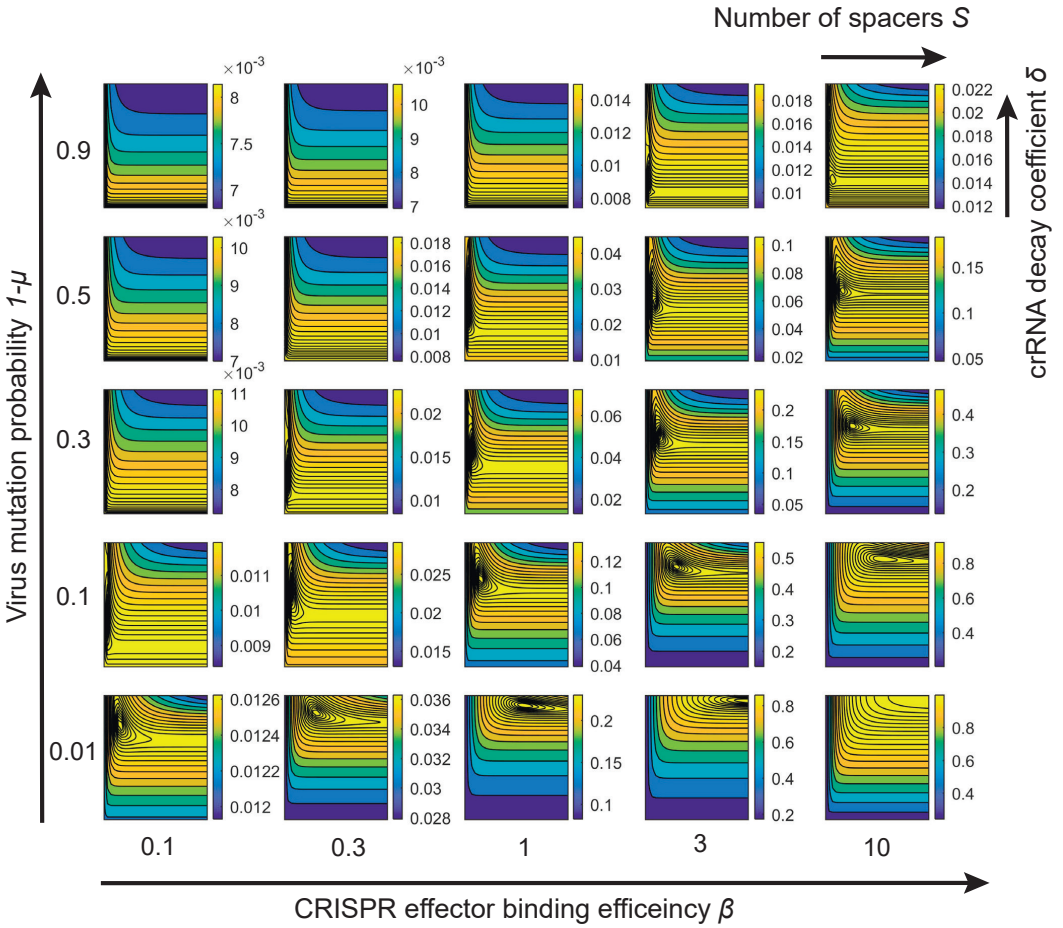

FIGURE 3.5: **Effects of mutation rate and binding efficiency.** A set of 25 panels illustrating how the survival probability depends on S and δ for various values of protospacer mutation probability $1-\mu$ and binding efficiency of effectors β. The δ and S axes in each small panel have the same range as in the panel A in Fig. 3.4, 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 $rNt = 5$.

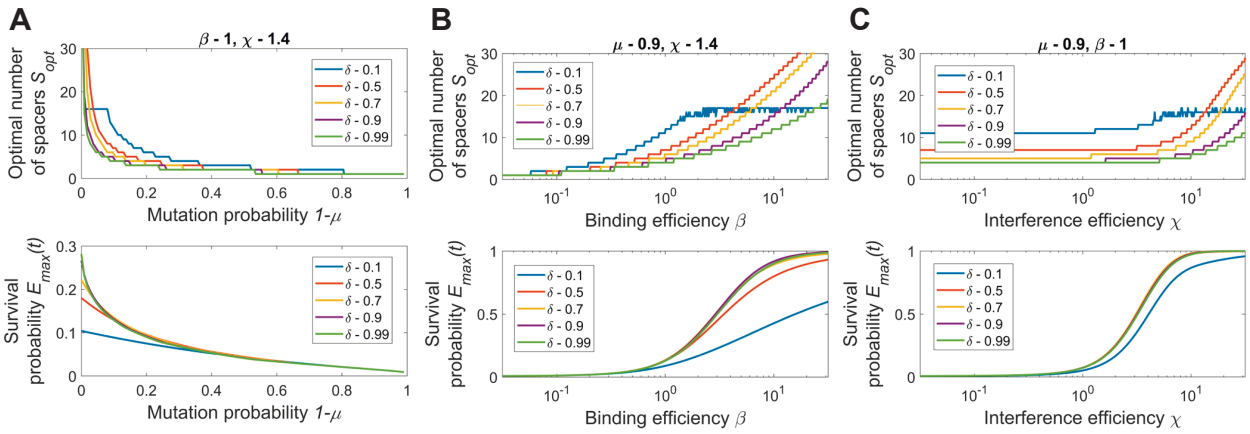

FIGURE 3.6: **Effect of parameters on the optimal number of spacers and the maximal survival probability.** The optimal number of spacers and corresponding survival probability as functions of one of the array-unrelated parameters: (A) As function of mutation probability $1-\mu$, other parameters are $\beta = 1$ and $\chi = 1.4$. (B) As function of binding efficiency β, other parameters are $\mu = 0.9$ and $\chi = 1.4$. (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.

On the other hand, a high binding β or interference efficiency χ lead to arrays with more spacers and higher survival probability (Fig. 3.6B, 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 the 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. 3.7, one sees that the parameters that produce high survival probability indeed correspond to arrays with relatively many spacers.

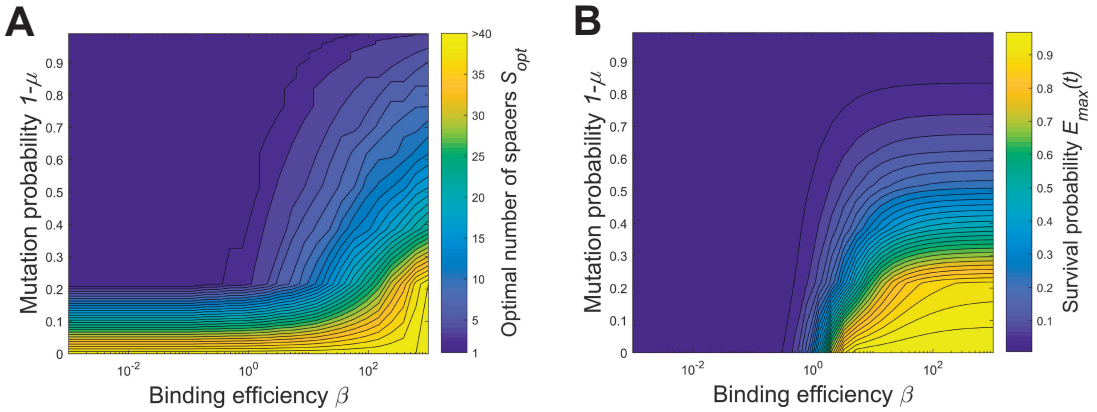

FIGURE 3.7: **The optimal number of spacers and maximal cell survival probability.** The optimal number of spacers (A) and the maximal cell survival probability (B) are shown vs. a range of binding efficiencies β and mutation probabilities $1-\mu$ for $rNt = 5$ and $\chi = 1.4$.

Figs. 3.6 and 3.7 lead to the 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 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.

3.3.3 Application: CRISPR-induced reduction in the viral burst

In the previous section, we estimated the number of CRISPR spacers that maximizes survival of a host cell. Here, we compute the number of spacers which minimize the viral burst (and thus the number of secondary infections) from a doomed host cell with still functioning CRISPR system. As in eq. (3.7), the total interference rate is assumed to be proportional to the total binding probability multiplied by the copy number of viral DNA. This is an overestimating approximation as in reality there is a spreading of a fixed number CRISPR effectors over increasing number of copies of viral DNA, which inevitably makes binding to any given protospacer less probable. Such a reduction in binding efficiency makes survival of viral DNA a "runaway" process: it becomes progressively less plausible to completely exterminate viral DNA after the first round of DNA replication. We also approximate viral DNA replication as a continuous process and obtain the following kinetic equation for the copy number of viral DNA $V(t)$,

$$\frac{dV(t)}{dt} = V(t)\left(D - a\sum_i B_i\right), \tag{3.23}$$

with the solution

$$V(t) = \exp\left[\left(D - a\sum_i B_i\right)t\right]. \tag{3.24}$$

Here D is the viral duplication rate and it is assumed that initially, the host cell contained a single copy of viral DNA, $V(0) = 1$

Without active CRISPR system, the number of viral DNA copies reaches the native burst size V_b after time θ,

$$V_b = \exp\left[D\theta\right]. \tag{3.25}$$

Assuming that the viral maturation time θ is not affected by CRISPR activity, the viral burst in the presence of CRISPR V_{CRISPR} becomes,

$$V_{CRISPR} = V_b \exp\left[-a\sum_i B_i\theta\right] = V_b \exp\left[-\chi\frac{\theta}{\tau}\sum_i B_i\right]. \tag{3.26}$$

The factor $\nu \equiv \theta/\tau$ is the number of cycles of replication of viral DNA and can be estimated from the burst size, $2^\nu = V_b$.

Steps analogous to those leading to eqs. (3.18) to (3.20) show that the burst size in host cells infected with viruses with S protospacers each having probability m_i to remain mutation-free is

$$V_{CRISPR} = V_b \prod_{i=1}^{S}\left\{1 - m_i\left[1 - \exp\left(-\nu\chi B_i\right)\right]\right\} \tag{3.27}$$

.

Comparing eq. (3.27) to eq. (3.21) reveals that the minimum of the product

$$\prod_{i=1}^{S}\left\{1 - m_i\left[1 - \exp\left(-\nu\chi B_i\right)\right]\right\} \tag{3.28}$$

maximizes the host cell survival probability when $\nu = 1$ and minimizes the viral burst size when ν equals to the number of cycles of replication of viral DNA. For a typical burst size $v_b = 100$, the number of replication cycles $\nu \approx 6.65$, which, as seen comparing left and right panels of Fig. 3.8, usually increases the optimal number of spacers (see also fig. 3.6(C) showing the

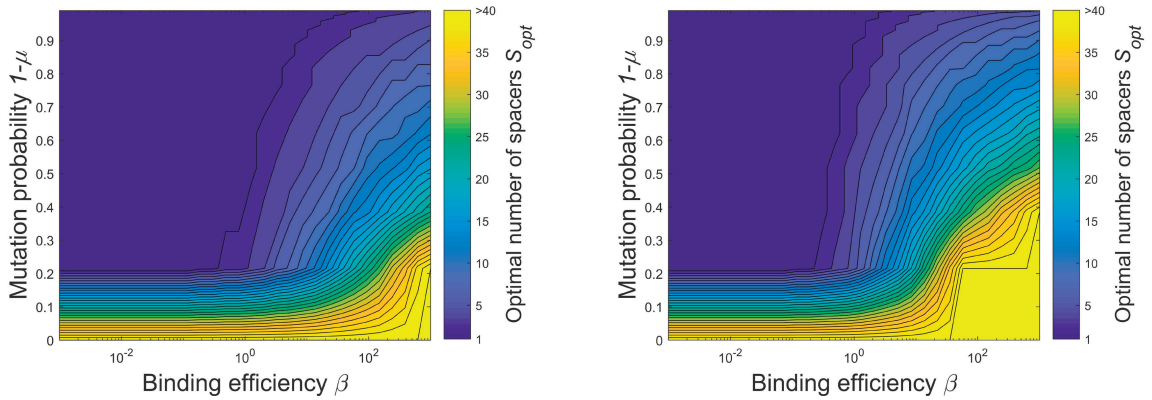

FIGURE 3.8: **Comparison of the optimal number of spacers for maximal cell survival probability and for viral burst reduction.** The number of spacers S that maximizes expression 3.28 for: the host cell survival with $\chi = 1.4$ (left panel) and the size of viral burst with $\chi' = \nu\chi = 1.4 * 6.65 \approx 9.3$ (right panel). $m_i = \mu^{i-1/2}$.

dependence of the optimal number of spacers on χ.)

3.3.4 Application: 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. (3.12), 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 (3.12) 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,

$$E_c(t) = \exp\left[-rNt\sum_{j=1}^{\nu} v_j\left(\prod_{i\in\{S_j\}}\{1-m_i[1-\exp(-\chi B_i)]\}\right)\right]. \tag{3.29}$$

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 (3.20), 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 (3.29) describes the 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

$$P_c = \prod_{k=1}^{S} v_k, \tag{3.30}$$

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.

$$E(t) = \sum_c E_c(t) P_c. \tag{3.31}$$

3.3.5 Results: Multiple viral species

A typical plot of $E(t)$ is presented in Fig. 3.9. 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. 3.4: 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. 3.4, 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. 3.4A and 3.9), 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 [74], it is unlikely that the CRISPR efficiency can be significantly higher

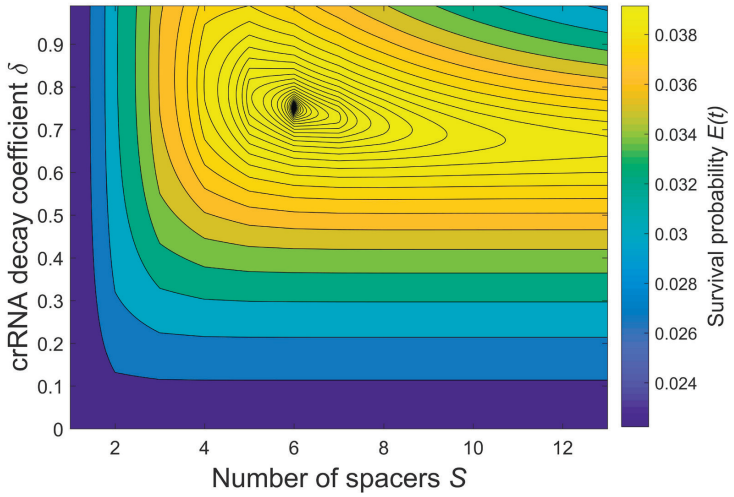

FIGURE 3.9: **CRISPR performance for two virus species.** Plot of the survival probability $E(t)$ as a function of crRNA decay coefficient δ and the number of spacers S of a cell confronting two different viruses with equal population sizes, $\nu_1 = \nu_2 = 0.5$. The binding efficiency is $\beta = 1$ and the interference efficiency is $\chi = 1.4$. Viral mutation probability $1 - \mu$ is equal to 0.1 and $rNt = 5$.

in vivo than the numbers shown in Fig. 3.9.

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. 3.10A). 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, the further increase in the number of spacers does not lead to any increase in the 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. 3.10B). 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.

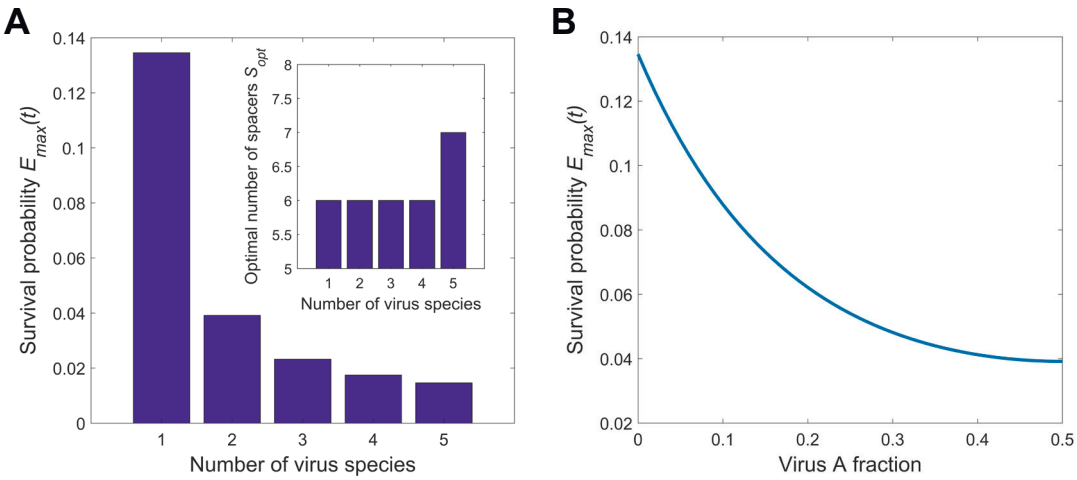

FIGURE 3.10: **Survival probability versus diversity of the predator virus pool.** Plots of the optimized over δ and S cell survival probability and the number of spacers vs the number of viral species and the composition of a two-virus pool for $\beta = 1$, $\chi = 1.4$, $\mu = 0.9$ and $rNt = 5$. (A) 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, $\nu_1 = ... = \nu_n = 1/n$. (B) The maximal survival probability vs the relative abundance of one of the viruses in a two-virus pool.

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

3.4 Discussion

The function of CRISPR-Cas as a 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 a 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 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 2.5 and 2.7 are applicable only to a particular ("average") set of virus-host coexistence scenarios, our more general estimates for the survival probability given in Eqs. (3.12,3.20) 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.

3.4.1 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 [80, 137]. While in our analysis we assumed that the viral environment (i.e. species composition and concentrations) is constant (except for the 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 and is reached by the time roughly estimated as the time interval between the subsequent spacer acquisition events 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-Cas 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 instance, if the host cell population evolve in the rapidly evolving viral environment, the model suggests that the number of spacers should be low. However, that could not be achieved without the evolution of spacer acquisition and loss rates which should be both high in this case.

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 systems are 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.

The last factor that contributes to the potential optimum is the difference between the newly acquired spacers and spacers that were selected and retained through the selection process. While the model is more applicable to

the new ones the selected ones undergo the complex process of selection and could be unique both from the perspective of time since acquisition and the probability to mutate. These selected spacers should be analyzed independently and could be a matter of future model extension within the provided framework.

3.4.2 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 [135, 136]. 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 [135]. 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 [80]. 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 a lower viral burst size and fewer secondary infections [74]. Our analysis based on [74] (S1 Appendix section 1) resulted in the estimate of the CRISPR failure probability

around 30% compared to 10^{-5} in [80].

3.4.3 Unequal crRNA abundance and importance of palindromic nature of CRISPR repeats.

One of the 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 the formation of hairpin by CRISPR repeats in the primary array transcripts [17, 153]. It is well known that hairpins have a potential to pause or terminate transcription elongation [154, 155]. 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 [14].

Another possible mechanism to control the abundance of crRNA derived from newer and older spacers is the binding of regulatory proteins that specifically target CRISPR repeats [156]. If these proteins act as transcription terminators, such binding also results in an exponential-like distribution of spacers. Also, it is possible that the spacer sequence itself holds the regulatory sequences such as terminator elements [150].

3.4.4 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 [157]. 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 [124], the cost of maintenance of *cas* genes [123], self-immunity [121] and blockage of beneficial horizontal gene transfer (HGT) [103]. 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 [125]. 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 [6, 43, 126]. For the cost of gene maintenance [123] and blockage of HGT [109], 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 systems did not affect the optimal number of spacers in the 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 the separation of the number of spacers question from the overall fitness. The factors described in this work affect the optimum number of spacers in the CRISPR array and the total fitness benefit of the 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 [106].

3.4.5 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 [61]. 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 cases with 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 the 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.

3.4.6 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 [115, 140], reducing the virus burst size (number of progeny viruses) [60, 74]. This is also related to the herd immunity - the effect that overall population can resist and limit the infection epidemic while each individual cells could die [158, 159]. This constitutes the second source of selection pressure on the CRISPR functioning.

We analyzed how to minimize the viral burst in section 1 of S1 Appendix. It appears that the condition for the minimum of the viral burst (S5) is similar to that for cell survival, (3.20), 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 (Figs. 5C and S1).

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.

3.4.7 Conclusions

- We theoretically predict the optimal number of spacers in a CRISPR array which falls into a reasonable range from the viewpoint of current experimental data and shows 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.

Chapter 4

Plasmid dynamics under a pressure of CRISPR-Cas

4.1 Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR associated protein) is a prokaryotic adaptive immune system. It protects cells from viral infections by recognizing and destroying viral DNA if it matches the record stored in cellular DNA in the form of spacers that compose a CRISPR array. Along with viruses, CRISPR-Cas systems also targets other foreign genetic elements such as plasmids. An active CRISPR-Cas system has a potential to eliminate plasmids from the host cell or even prevent a host cell from acquiring plasmids at all.

However, as in the case with viruses, a CRISPR-Cas system does not provide a guaranteed protection against plasmids. Plasmids can escape elimination (also called interference) by CRISPR-Cas system in several ways: A plasmid can avoid recognition due to a mutation either in the DNA segment targeted by CRISPR-Cas system (a protospacer) or in the targeting its spacer in cellular DNA, and through a disruptive mutation or knockout of the CRISPR-Cas system altogether [106]. Also, a fast enough plasmid replication can kinetically outcompete plasmid degradation by the CRISPR-Cas

system when the match between the spacer and its protospacer is not perfect [60]. In this case, the action of the CRISPR-Cas system does not eliminate all plasmids but only reduces the equilibrium population of plasmids below its native level.

To study how successful the CRISPR-Cas plasmid interference is, we perform direct experiments, transforming plasmids into modified *E. coli* cells with induced CRISPR-Cas genes. We observe that plasmids survive in some cells with the active CRISPR-Cas system. While only a small fraction (less than 1%) of cells retain plasmids under the pressure of CRISPR-Cas system, plasmids remain in those cells and their descendants for many generations, so that such transformed cells form colonies. None of the known mechanisms outlined above explains the observed plasmid survival: Additional experiments confirm that the CRISPR-Cas system in such cells is active and the match between plasmid protospacer and the CRISPR-Cas spacer remains perfect. Extinction of plasmids in the majority of cells indicates that the plasmid degradation rate generally exceeds the replication rate, which rules out the simple kinetic explanation as well.

To understand such a surprising outcome of the interaction between plasmids and CRISPR-Cas system, we took a closer look at the kinetics of plasmid replication and interference. It turns out that under rather general assumptions about the dependences of plasmid replication and interference kinetics on plasmid copy number, there is a possibility that the replication rate exceeds the interference rate for certain intermediate numbers of plasmids per cell. At the same time, for one or a few cellular plasmids, the interference outcompetes replication, which explains the extinction of plasmids in all but a small fraction of cells. The cells that retain plasmids do so by a pure chance when a sequence of random plasmid replication and interference events starts and then predominantly consists of the former ones. To

make this explanation more quantitative, we developed a model of plasmid population dynamics that takes into account stochastic plasmid replication and CRISPR-Cas interference events and the plasmid redistribution during cell division. The model shows that there indeed exists a range of rate constants that lead to the selective effect: Initially small plasmid population (usually just a single plasmid per cell) has a very high probability to go extinct, while if plasmids survive and replicate to reach a sufficiently high number, their population is maintained almost indefinitely. The CRISPR-Cas activity drives the cellular population to split into subpopulations with and without plasmids, creating a bimodality in cellular type distribution.

These explanations and predictions are accompanied by several follow-up experiments. First, when the CRISPR-Cas system is activated in a cell with already well-established plasmid population, the probability for a cell to lose all plasmids in a long-term is much lower than the probability to do so for a cell with initially one plasmid. In this scenario when the initial plasmid number is large, the plasmid population is gradually reduced by the CRISPR-Cas system to a new steady state without passing through the low copy number bottleneck. Second, re-plating experiments show that the plasmid number in cells that retain plasmids converges to the same average, which is independent of the initial number of plasmids at the moment of CRISPR-Cas activation. At the same time, the number of cells that retain plasmids varies by orders of magnitude depending on the initial (when CRISPR-Cas system is activated) number of plasmids in a cell, which confirms the relevance of stochastic nature of plasmid survival at low copy number. In addition to explaning a puzzling experimental founding and offering an indirect way to assess the CRISPR-Cas interference kinetics, we speculate on possible evolutionary advantages of such "imperfectly tuned" CRISPR-Cas systems, which allow a small fraction of cells to retain potentially beneficial plasmids.

4.2 Methods and Models

4.2.1 Strains and plasmids

We used E. coli KD263 (K-12 F+, lacUV5-cas3 araBp8-cse 1, CRISPR I: repeat-spacer g8-repeat, CRISPR II deleted) cells as described in [160]. The plasmid pG8 carrying a 209-bp M13 fragment with the g8 protospacer has been constructed from the pT7blue plasmid as described in [58], and a pRSF-G8 plasmid containing also g8 protospacer has been constructed from a pRSF plasmid as described in [61].

4.2.2 CRISPR interference assays

To assay for plasmid interference, 50 µl of overnight culture was diluted with 5 mL of Luria-Bertani (LB) broth. The diluted culture was incubated in LB medium at 37° C in the presence (CRISPR ON) or in the absence (CRISPR OFF) of 1 mM arabinose and 1 mM IPTG until the culture OD_{600} reached 0.6. The electrocompetent cells were prepared using a standard protocol [161] and transformed with 5 ng of plasmids containing protospacers. Next, the transformants were put in tubes containing 1 mL Lb with 1 mM arabinose and 1 mM IPTG for CRISPR ON cultures and 1 mL LB for CRISPR OFF cultures. After 1-h 37° C outgrowth, 50 µl aliquots of serial dilutions of transformation mixtures were plated onto LB agar plates containing 100 µg/ml ampicillin (Ap) or 50 µg/ml kanamycin (Km) with (CRISPR ON) or without (CRISPR OFF) inducers. Plates were incubated at 37° C overnight. The efficiency of transformation (EOT) was determined as a number of colony-forming units (CFU) per 1 µg of plasmid DNA (Fig. 4.3). These procedures were repeated 3 times for each of the two plasmids.

To test the condition of plasmids in CRISPR ON transformants, the plasmid from ten randomly chosen such colonies were isolated using GeneJET

Plasmid Miniprep Kit (Thermo scientific) and retransformation into new CRISPR ON and CRISPR OFF cells were carried out (Fig. 4.5B).

To determine if the CRISPR-Cas system remained active in the CRISPR ON transformants, retransformation of ten randomly chosen individual colonies was performed with the complementary plasmid also containing g8 protospacer: The cells initially transformed with pG8 plasmid received pRSF-G8 and vice versa, Fig. 4.5C. The efficiency of retransformations was evaluated as described above.

4.2.3 Real-time PCR assay of plasmids

The real-time polymerase chain reaction was used to estimate the difference in a plasmid copy number between transformants growing with and without inducers. Five colonies from each plate were assayed with qPCR. Each qPCR reaction with two groups of oligonucleotides for gyrase and β-lactamase genes was carried out for each of 3 technical repeats in a 20 µl reaction volume with 0.8 units of HS Taq DNA polymerase (Evrogen) and 0.01 µl of Syto13 intercalating dye (LifeTechnology) using DTlite4 (DNA-Technology) amplifier. The results of qPCR with plasmid-specific primers were normalized to genomic DNA. The primers for pT7blue-g8rev were: Bla_dir TGAGTATTCAACATTTCCGTGTCG, Bla_rev CGAAAACTCTCAAGGATCTTACCG; for pRSF-g8rev: pRSF_ori_dir GTCCGCTCTCCTGTTCCG, pRSF_ori_rev AGCCTGAGCTATGAC AGCG. For genomic DNA we applied the primers GyrA_dir CGGTCAACATTGAGGAAGAGC and GyrA_rev TACGTCACCAACGACACGG. This procedure of real-time PCR was repeated at least three times for the transformants and replated colonies.

To improve the precision of our measurement of the per cell plasmids copy number (PCN), we calibrated the primers as in [162]. For this aim we

took individual colonies of the transformants and resuspended them in a series of samples, diluting each subsequent sample 10 times. For real-time PCR we assayed five serial dilutions, doing three technical repeats for each dilution. We used only the results of real-time PCR with the deviation less than $0.1\Delta Ct$ among technical repeats of one dilution. The efficiency of primers was calculated from the average slope of the plot of the logarithm of the concentration of dilutions vs. Ct. We performed such calibration for each pair of primers three times and obtained that the average amplification factor of the primers Bla_dir and Bla_rev were 2.0 (amplification efficiency 100%), of primers GyrA_dir and GyrA_rev were 1.9 (amplification efficiency 90%) and of primers pRSF_ori_dir and pRSF_ori_rev were 2.1 (amplification efficiency 110%). Using the efficiency of primers and the results of real-time PCR, we estimated PCN.

4.2.4 Replating of transformants

Randomly chosen individual colonies of the CRISPR ON and CRISPR OFF transformants were replated on three types of selective media: LB with antibiotic (Ab) and inducers (Ind) to maintain the CRISPR-Cas activity, LB with Ab (to determine the number of plasmid-bearing cells) and LB (to determine the total number of cells) (Fig. 4.8). To do so, each colony was diluted in 200 µl LB, and 5 µl of culture was plated on the media in steps of 4-fold dilutions. The CFU were counted on each plate. The colonies from plates with Ab/Ind were used for the second replated. Each replating was repeated at least 3 times.

4.2.5 Dynamics of replication and degradation of plasmids

The dynamics of plasmid replication can be quite complex, yet it has two universal limits: For a few plasmids, the replication rate is proportional to the

number of plasmids (i.e. the replication rate per plasmid is constant), while for the target (standard) concentration of plasmids $[Pl]_{st}$, the replication rate is zero. As often done [163], we approximate such dynamics by the logistic model,

$$\left.\frac{d[Pl]}{dt}\right|_{replication} = k_d[Pl]\left(1 - \frac{[Pl]}{[Pl]_{st}}\right). \tag{4.1}$$

The coefficient k_d is the per capita rate of plasmid replication in the low concentration limit. The sign $[x]$ indicates the concentration of a substance x. Assuming that the volume of a cell stays approximately constant, we define a concentration as the number of molecules per cell, and in the following we use the terms "concentration" and "copy number" interchangeably.

As a catalytic process, the interaction of CRISPR-Cas complexes Cr with plasmids Pl,

$$Pl + Cr \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} PlCr \overset{k_2}{\to} Cr + \emptyset, \tag{4.2}$$

is assumed to be well-described by the Michaeles-Menten kinetics,

$$\left.\frac{d[Pl]}{dt}\right|_{cutting} = -k_2[PlCr] = \tag{4.3}$$
$$-k_2\frac{[Pl]_0 + [Cr]_0 + \chi - \sqrt{([Pl]_0 + [Cr]_0 + \chi)^2 - 4[Pl]_0[Cr]_0}}{2}.$$

Here, as in the standard Michaeles-Menten derivation, the stationarity of concentration of the CRISPR-Cas-plasmid complex is assumed, the generalized dissociation constant χ is defined as

$$\chi \equiv \frac{k_1^- + k_2}{k_1^+}, \tag{4.4}$$

and no assumption is made on overabundance of the catalyst (CRISPR-Cas) or the substrate (plasmid). The total (bound in the $PlCr$ complex and free)

concentrations of plasmids and CRISPR-Cas complexes are $[Pl]_0$ and $[Cr]_0$.

Assuming that replication only increases the plasmid concentration, or in other words, $[Pl]$ in (4.1) never exceeds $[Pl]_{st}$, we define a one-step birth-death process [164] for the population of plasmids. The probabilities of increasing or decreasing the population of plasmids by one $\beta_{[Pl]}$ and $\delta_{[Pl]}$ are given by $d[Pl]/dt|_{replication}$ (4.1) and $d[Pl]/dt|_{cutting}$ (4.3). The master equation that describes the temporal evolution of probability $P_{[Pl]}(t)$ to find a cell having $[Pl]$ plasmids at time t [164] reads

$$\frac{dP_{[Pl]}(t)}{dt} = \beta_{[Pl]-1}P_{[Pl]-1}(t) + \delta_{[Pl]+1}P_{[Pl]+1}(t) - (\beta_{[Pl]} + \delta_{[Pl]})P_{[Pl]}(t). \quad (4.5)$$

4.2.6 Redistribution of plasmids during cell division

In addition to cutting and replication of plasmids, the per cell plasmid copy number is also affected by cell division, which happens every $\tau \approx 20$ min. A conservative estimate would be that the redistribution of plasmids between daughter cells is completely random (in reality it is biased towards equal or half and half distribution). Assuming also that the act of cell division happens fast (instantaneous) compared to the replication and cutting of plasmids, the outcome of the redistribution process can be described by the binomial distribution with the probability for each plasmid to go into any of two daughter cells equal to $1/2$. If a cell before the division had j plasmids, the probability B_{ij} to find $0 \leq i \leq j$ plasmids in one of the daughter cells is

$$B_{ij} = \frac{j!}{i!\,(j-i)!}\left(\frac{1}{2}\right)^j. \quad (4.6)$$

4.2.7 Simulation procedure

As presented above, the temporal dynamics of plasmid copy number in a cell is approximated by a sequence of periods of continuous evolution, described

by the master equation (4.5), each followed by the instantaneous redistribution between daughter cells, described by the binomial distribution (4.6). To estimate the distribution of plasmids in cells in CRISPR ON colonies after several hours of growth, we implement the following numerical procedure (see fig. 4.1):

- For a given set of plasmid replication and CRISPR interference parameters k_d, $[Pl]_{st}$, k_2, χ, and $[Cr]_0$ (see table 4.1), we tabulate the replication and cutting rates $\beta_{[Pl]}$ and $\delta_{[Pl]}$ for all possible plasmid copy numbers, $1 \leq [Pl] \leq [Pl]_{st}$.

- We numerically integrate the master equation (4.5) till the cell cycle time τ, starting from every possible initial number of plasmids j, $0 \leq j \leq [Pl]_{st}$. Naturally, the solution with zero initial plasmids will always be zero plasmids with probability one.

- The probabilities C_{ij} for a cell to end up with i plasmids at time τ after starting with j plasmids at $t = 0$,

$$C_{ij} \equiv P_i(\tau),\ P_k(0) = \delta_{k,j}, \tag{4.7}$$

  are collected into the matrix $\hat{C}$. Another matrix $\hat{B}$ is composed of binomial probabilities B_{ij} (4.6).

- The probability to find k plasmids after time t is given by the $k+1$th element of the $[Pl]_{st}+1$-dimensional vector $\vec{P}$,

$$\vec{P} = \hat{C}\left(\hat{B}\hat{C}\right)^N \vec{P}(0), \tag{4.8}$$

  where N (equal to the integer part of t/τ) is the number of cell cycles and the initial condition $\vec{P}(0)^T$ indicates how many plasmids were

in each cell when the CRISPR-Cas system was activated. Here we assumed that the number of plasmids in a cell is assessed at the final stage of cell cycle just before cell division, thus an extra multiplication by $\hat{C}$. Alternatively, when the number of generation is not very large, this probability can be computed more efficiently by direct solution of the master equation (4.5) for periods of time between cell division alternated with binomial redistribution of plasmids between daughter cells according to (4.6). In such case, we do not need to compute the matrix $\hat{C}$).

TABLE 4.1: List of parameters used in the model of the plasmid dynamics under a pressure of CRISPR-Cas

Model parameter	**Parameter name**	**Parameter description**
$[Cr]_0$	Number of CRISPR-Cas complexes	Total number (bound and unbound) of CRISPR-Cas complexes in the cell.
$[Pl]_{st}$	Plasmid standard copy number	The plasmid copy in the cell that is reached without CRISPR-Cas activity - maximal plasmid copy number in the cell.
k_d	Plasmid replication rate	Per capita rate of plasmid replication in the low concentration limit.
χ	CRISPR-Cas complex dissociation constant	Dissociation constant of the establishing of CRISPR-Cas-plasmid complex.
k_2	CRISPR-Cas catalytic rate constant	Plasmid degradation by CRISPR-Cas rate constant.

The evolution of the probability density $P_k(t)$ for the replication and interference rates (4.1) and (4.3) plotted in Fig. 4.7A is shown in Fig. 4.7B for cells initially having 1 plasmid, ($P_k(0) = \delta_{k,1}$ being the typical initial condition in a CRISPR-ON experiment) and Fig. 4.7C for cells initially having the target number of plasmids, ($P_k(0) = \delta_{k,[Pl]_{st}}$ being the initial condition for replating

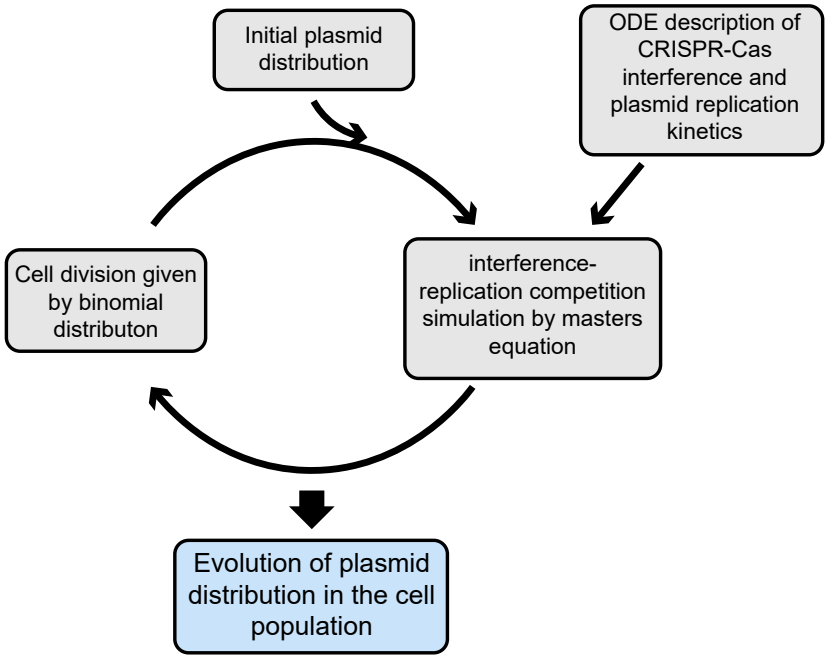

FIGURE 4.1: **Plasmid dynamics simulation scheme** Based on the kinetics of plasmid replication and interference a master equation for plasmid dynamics was constructed. Initial plasmid distribution is fed into the cycles of cell population generation simulations consisting of growth step simulated by masters equation and division step calculated as a binomial distribution of plasmids between daughter cells.

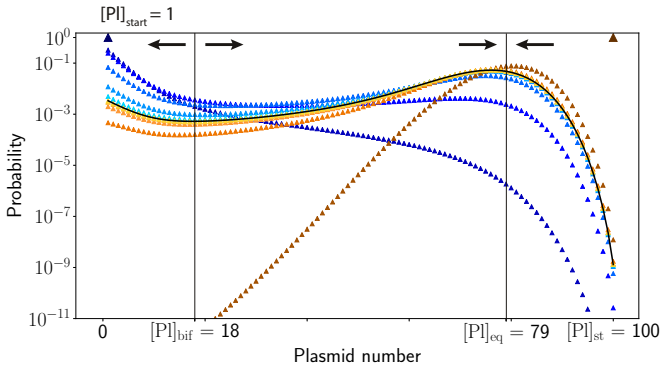

FIGURE 4.2: **Convergence of plasmid number probability distribution to the universal scaling form** The blue triangles show the evolution of $P_k(t)$ when cells initially had a single plasmid, while the orange/brown squares show the evolution of $P_k(t)$ when cells initially had $[Pl]_{st}$ plasmids. Shades of blue and red correspond to the different generations of cells from generation 0 (dark triangle and square) to generation 5 (the lightest blue and orange). Both families of curves converge to the universal asymptotic curve shown by a black line.

the CRISPR OFF cells on plates with inductor). The plots in Fig. 4.7 were computed using the following parameters $k_d = 0.3$, $[Pl]_{st} = 100$, $k_2 = 0.5$, $\chi = 1$, and $[Cr]_0 = 10$.

As most birth-death processes, this stochastic process of plasmid replication, cutting, and redistribution has the unique convergent steady state $\vec{P}(\infty)^T = (1, 0, \ldots, 0)$, corresponding to the extinction of all plasmids. However, after a few cell cycles, while the component $P_0(t)$ that corresponds to the fraction of cells with no plasmids steadily grows, other components $P_k(t)$, $k = 1, \ldots, k_{st}$ that correspond to the probability to have a non-zero number of plasmids approach a steady state scaling form,

$$P_k(t) = f(t)\tilde{P}_k, \; k = 1, \ldots, [Pl]_{st}, \tag{4.9}$$

shown in Fig. 4.2. The slowly-decaying function $f(t)$ represents a universal convergence to the absorbing state $\vec{P}(\infty)^T = (1, 0, \ldots, 0)$.

4.3 Results

4.3.1 Survival of plasmids in cells with active CRISPR

The KD263 *E. coli* cells contain inducible *cas* genes and a miniature CRISPR array with a single g8 spacer [160]. Comparing the transformation efficiency of induced (CRISPR ON) and uninduced (CRISPR OFF) KD263 cells with a plasmid containing the G8 protospacer and a functional ATG PAM allows one to detect CRISPR interference. After the transformation, CRISPR OFF cells are plated on plates supplemented with antibiotic only. Transformed CRISPR ON cells are plated on a medium that contains both antibiotic and inducers of *cas* genes expression (fig. 4.3). Compared to CRISPR OFF KD263, there is approximately 200 times less ampicillin-resistant transformants formed after the same amount of pG8 plasmid is transformed in induced, CRISPR ON cells. For another plasmid pRSF-G8, there are approximately 40 times less kanamycin-resistant colonies on CRISPR ON medium than CRISPR OFF one, (fig. 4.5A). In both cases, the antibiotic-resistance colonies formed by CRISPR ON cells appear healthy and indistinguishable from CRISPR OFF cell colonies.

A question arises as to the nature of CRISPR ON transformants. Likely explanations could be inactivation of CRISPR-Cas system in cells forming antibiotic resistance colonies or the presence of plasmids with mutated protospacer in transformed cells. To test for the latter possibility, we performed an experiment involved purification of plasmid from ten randomly chosen individual colonies of CRISPR ON and retransformation in CRISPR ON and CRISPR OFF cells. In every case, a 200 drop in transformation efficiency in induced cells was observed. We, therefore, conclude that plasmids present in CRISPR ON cells are subject to interference by CRISPR effector charged with crRNA with g8 spacer. To determine whether CRISPR ON cells forming colonies on selective medium contains a functional CRISPR-Cas system,

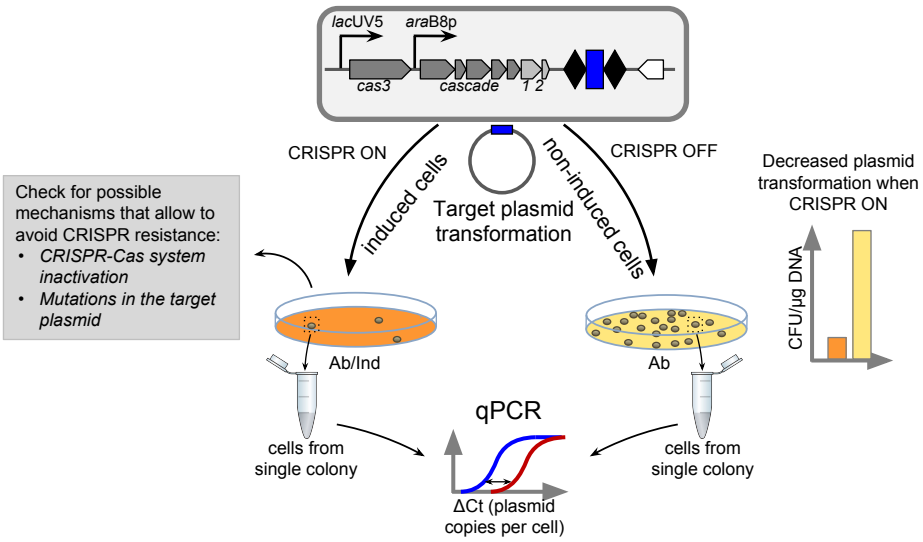

FIGURE 4.3: **Scheme of the initial plating experiments.** The strain *E.coli* K12 KD263, expressing *cas* genes from the inducible promoter and containing CRISPR array with a single g8 spacer, and the plasmids pG8 and pRSF-G8 bearing g8 protospacer were used as a testing model. The induced (CRISPR ON) and non-induced (CRISPR OFF) KD263 cells were prepared and transformed with ampicillin-resistant (ApR) pG8 plasmid or kanamycin-resistant (KanR) pRSF plasmid. The induced transformants were plated on medium with antibiotic and inducers (Ab/Ind), while the non-induced ones were plated on antibiotic only (Ab). To compare plasmid populations in colonies from Ab/Ind and Ab plates, the real-time PCR assay of random individual colonies was performed.

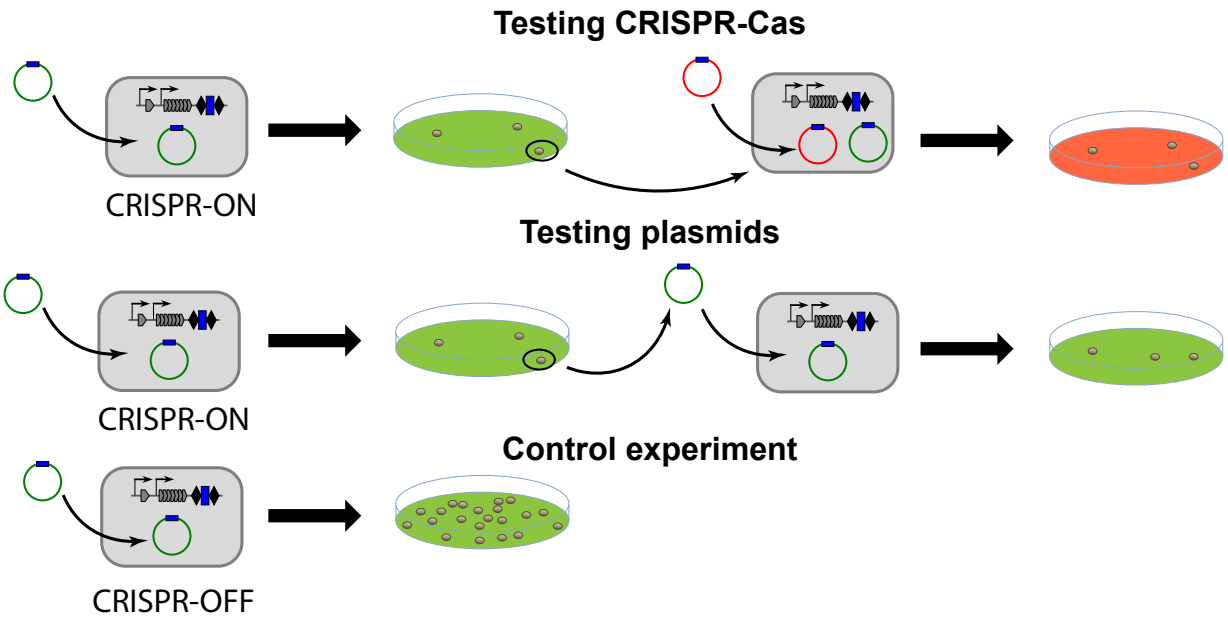

FIGURE 4.4: **Shcematic depiction of experiments, testing CRISPR-plasmid interference.** Plasmids holding antibiotics resistance were transformed in the host cells with CRISPR-Cas systems targeting these plasmids and plated on the antibiotics medium. Control experiments with inactivated CRISPR-Cas (CRISPR-OFF) showed a high number of colonies. Activated CRISPR-Cas (CRISPR-ON) showed a low number of colonies. To test CRISPR-Cas the survived cells were transformed with other plasmids that contain different antibiotic resistance but same protospacer. To test that target plasmid protospacer remained un-mutated plasmids from survived cells were extracted and transformed into the original host strain.

competent cells were prepared from individual CRISPR ON and CRISPR OFF transformed colonies. In each case, cultures used to obtain competent cells were grown either with or without *cas* gene inducers and transformed with pRSF plasmid, carrying the G8 protospacer with a functional PAM. CRISPR interference was determined by comparing transformation efficiency into induced and uninduced cells. The results, presented in fig. 4.5B, show that cells derived from pG8 transformed CRISPR ON colonies interfered with pRSF-G8 transformation as efficiently as the CRISPR OFF control cells. The same situation was observed for pRSF transformed CRISPR ON colonies fig. 4.5C. We, therefore, conclude that colonies formed by CRISPR ON cells carrying a plasmid with protospacer matching crRNA spacer are formed by cells with active CRISPR-Cas system (see fig. 4.4).

4.3.2 Qualitative explanation of plasmid survival

To explain the intriguing effect of successful transformation in a small fraction of CRISPR ON cells and the apparent flexible and history-dependent response of plasmid population to CRISPR-Cas interference, we take a closer look at the dynamics of the plasmid population in a cell. The plasmids populations are formed by two competing mechanisms, plasmid interference (cutting) by CRISPR-Cas complexes and plasmid replication. Our observation that almost all cells become plasmid-free indicates that the former process is usually faster. However, the existence of cells with surviving plasmids suggests that there is a small probability of plasmid replication successfully outcompeting the interference.

It is fair to assume that for a single or a few plasmids when the plasmid concentration is the limiting factor in cutting and replication kinetics, the per plasmid rates of both those processes are constant, i.e. independent of the plasmid copy number. It means that the per population rates of both

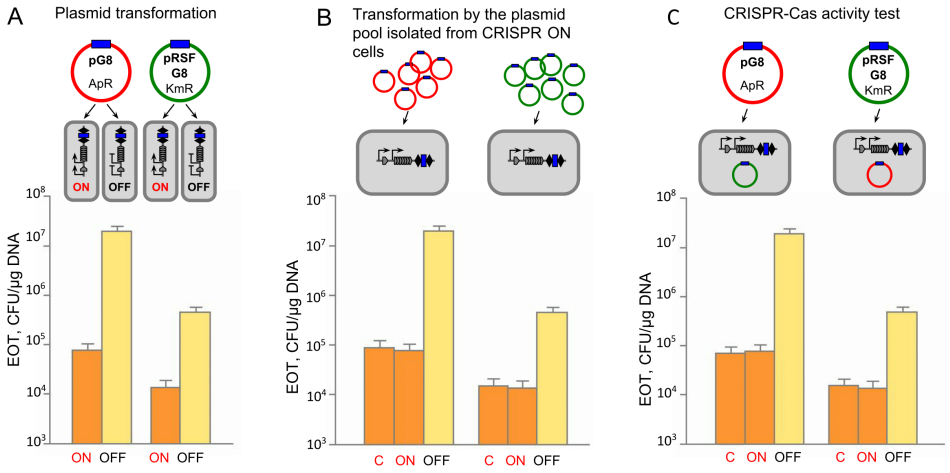

FIGURE 4.5: **Results of transformation and testing of the transformants.** Two different plasmid, pG8 and pRSF-G8, with g8 protospacer are used to transform induced (CRISPR ON) and non-induced (CRISPR OFF) KD263 cells (see fig. 4.3). The efficiency of transformation (EOT) is determined in CFU/µg plasmid DNA for both plasmids. The orange bars show EOT for CRISPR ON cells, the yellow bars show EOT for CRISPR OFF cells. (A) The number of colonies on plates with antibiotic and inducers is $\approx$ 200 times less than that on plates with antibiotic only for pG8 plasmid (first and second bars) and $\approx$ 40 times less for pRSF-G8 (third and fourth bars). (B) Test for escape mutations in the g8 protospacer. The pG8 and pRSF-G8 plasmids were purified from the induced transformed cells and retransformed into the original KD263 cells. The efficiency of transformation of retransformed plasmids (bars C) was the same as of the original plasmids (bars ON). (C) To test the CRISPR-Cas system of the transformants with pRSF-G8, the plasmid pG8 was used for retransformation of those transformants. The same cross-retransformation approach was implemented for the transformants with pG8 using the plasmid pRSF-G8. New transformations occurred with the same efficiency (bars C) as the initial ones (bars ON).

those processes increase linearly with the number of plasmids. Yet when the number of plasmids becomes large and comes close to the stationary number of plasmids in a cell ($[Pl]_{st} \approx 100$ in our experiments), the replication rate should approach zero. At the same time, for all reasonable forms of interference kinetics, an increase in the number of plasmids results in a progressively smaller increase in the interference rate, which finally saturates to a constant when the concentration of plasmids becomes much higher than that of CRISPR-Cas complexes. A comparison between the cutting and replication rates that satisfy those general constraints indicates that three scenarios are possible:

- The replication rate is always lower than the interference rate, fig. 4.6A.
- The replication rate is higher than the interference rate, fig. 4.6B.
- There is a window of plasmid number for which the replication rate exceeds the interference rate, while beyond this window the interference rate is higher, fig. 4.6C.

Evidently, the first scenario leads to a quick loss of plasmids in all cells, while the second scenario results in survival of plasmids in the majority of cells. However, the third option holds a potential explanation for the observed survival of plasmids in a small fraction of cells: Since all cells initially have just one plasmid, almost all of them lose plasmids since the cutting rate is larger than the replication rate for low plasmid copy number. However, due to an intrinsic stochasticity of cutting and replication events, there is a small but finite possibility that more replication than cutting events initially occur. If such a lucky for plasmids sequence of events takes the plasmid population over the bifurcation threshold, above which the replication rate exceeds the cutting rate (marked as $[Pl]_{bif}$ in fig. 4.6C), the plasmid population will likely survive and continue to expand from this point deterministically

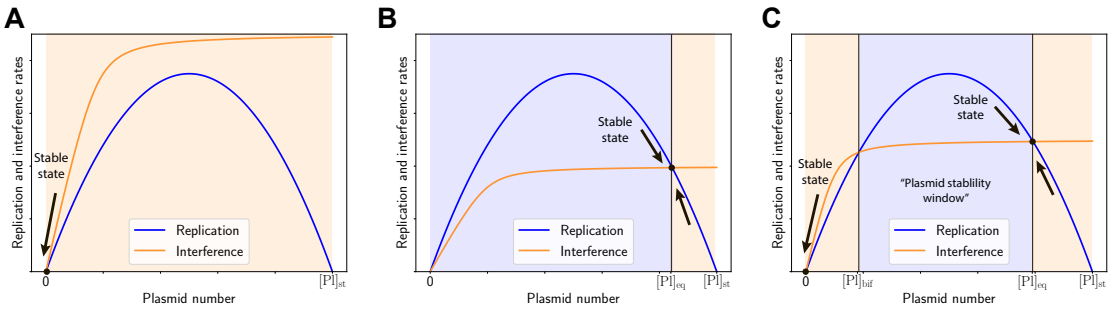

FIGURE 4.6: **Comparison between the plasmid interference and replication rates.** Three possible relations between plasmid replication (blue lines) and CRISPR-Cas interference (orange lines) rates. Intervals of plasmid copy number for which replication or interference dominate are shown by blue or orange shading. Stable equilibrium points for plasmid population are shown by black dots. (A) The interference rate is higher than the replication rate for any number of plasmids, so that the introduced plasmids quickly become extinct. (B) The replication rate is higher than the interference rate so that the plasmid copy number quickly reaches the equilibrium point $[Pl]_{eq}$. (C) There exists an intermediate range of plasmid copy numbers, $[Pl]_{bif} < [Pl] < [Pl]_{eq}$, where the replication rate is higher than the interference rate, while beyond this range the interference dominates.

until reaching $[Pl]_{eq}$. Clearly, a larger excess of the interference rate over the replication rate for small plasmid copy number and a higher threshold number of plasmids $[Pl]_{bif}$ reduce the probability of plasmid survival.

In the Methods and Models section, we outline a quantitative analysis of this survival scenario, which confirms that the reasoning presented above indeed explains the survival of plasmids. It is based on the numerical solution of a master equation, which describes the time evolution of the probability $P_n(t)$ for a cell to have n plasmids at time t. The master equation accounts for the plasmid replication and interference processes, which are assumed to follow the logistic dynamics and Michaelis-Menten kinetics and the binomial partition of plasmids between two daughter cells at cell division is treated as an instantaneous process. The results of the master equation solution are shown in panels B and C in fig. 4.7 for the initial number of plasmids equal to 1 and $[Pl]_{st}$.

As seen in fig. 4.7 that the competition between the interference and plasmid replication produces two cell subpopulations, one having a substantial number of plasmid distributed around $[Pl]_{eq}$, and the other completely devoid of plasmids. The probability for a cell to retain plasmids quickly drops in the first few generations and levels after 5-10 generations. Naturally, this probability strongly depends on the kinetics of plasmid replication and CRISPR-Cas interference and might be unique for each type of plasmid and CRISPR-Cas system.

It follows from our explanation that the fraction of cells that lose (and, reciprocally, retain) plasmids after the initial transitory period of 5-10 generations depends on the initial plasmid number in a cell with active CRISPR system. In terms of experimental scenarios, this fraction depends on whether one or a few plasmids are put into a cell with immediately activated (or already active) CRISPR-Cas system or the CRISPR-Cas system is turned on in cells with already pre-existing stationary plasmid population $[Pl]_{st}$. In

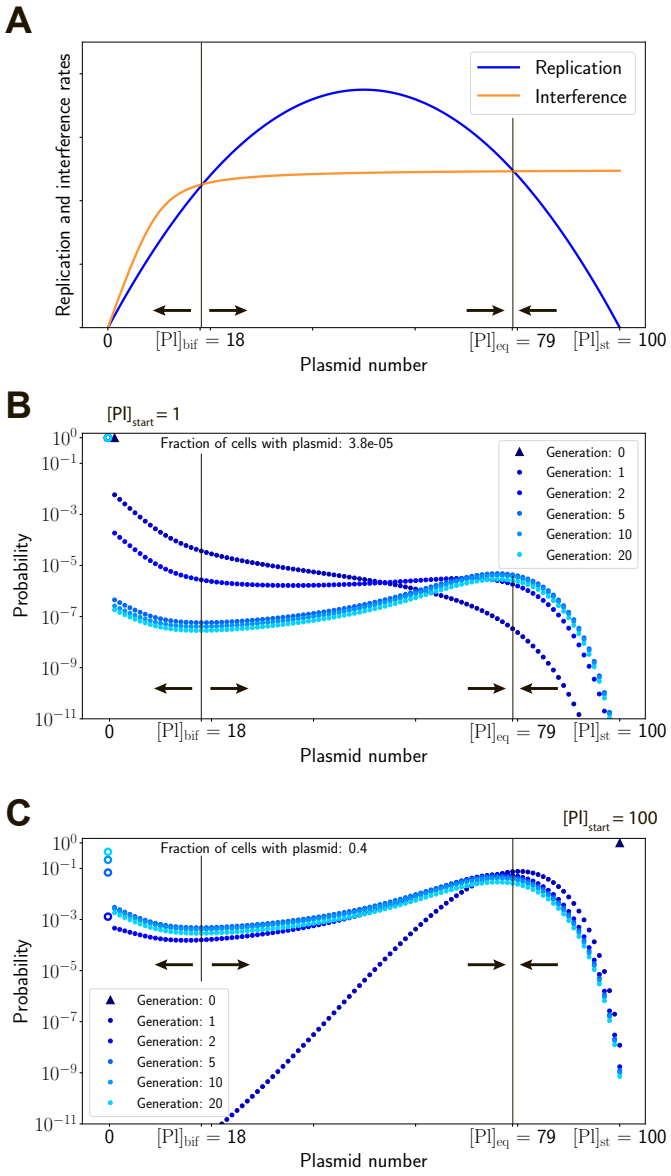

FIGURE 4.7: **The probability for a cell to have n plasmids after a given number of generations** (A) The replication (blue line) and interference (orange line) rates used in the solution of the master equation (4.5) are parametrized as the logistic (4.1) and Michaelis-Menten (4.3) kinetics. Vertical black lines in all panels show stable and unstable fixed points of plasmid dynamics, their stability is shown by converging and diverging arrows. (B) The probability $P_n(t)$ for a cell to have n plasmids after the 1-st, 2-nd, 5-th, 10-th, and 20-th generations are shown by dots of varying shades of blue. For each number of generations, the probability is computed for the time just before the partition of plasmids between two daughter cells. Initially, a single plasmid was introduced into a cell, which is marked by a triangle in the upper left corner. Empty circles, also marked by shades of blue of the corresponding generation, show the fraction of cells that lost all plasmids. (C) Same as in panel B, but for the initial number of plasmids equal to $[Pl]_{st}$, marked by a triangle in the upper right corner. The parameters used in this solution are listed in the Methods section. Comparing panels (B) and (C) it is visible that the probability distribution for a cell to have n plasmids after ≈ 10 generations converges to the universal form, shown by a light blue line in both panels. However, the fractions of cells with plasmids is much larger for multiple initial plasmids (C) than the single one (B), hence the light blue line in (C) is higher than that in (B).

the first case, we expect that majority of cells will lose plasmids, while in the second case a substantial fraction of cells will retain plasmids. For our experimental parameters, the fraction of cells that retain plasmids after 20 generations is 4×10^{-5} when the single plasmid is transformed into CRISPR ON cells and 0.4 when CRISPR–Cas is turned ON in cells with equilibrium plasmid population.

Nevertheless, the distribution of plasmids in cells that retain plasmids converges after ~ 10 generations to the universal form which does not depend on the initial number of plasmids and is determined by the kinetics of interference and replication. The universal distribution is shown by the light blue lines in panel B of fig. 4.7 for a single initial plasmid per cell and in panel C for $[Pl]_{st} = 100$ plasmids per cell. Since the average number of plasmids per cell that retained plasmids converges after several generations to a rather large number $[Pl]_{eq} \gg 1$ (evidently, this number is also universal and independent of the initial number of per cell plasmids) the subsequent probability to lose all plasmids becomes quite low, and such cells form colonies that survive indefinitely.

These analyses lead to three main hypothesis:

- The distribution of the plasmids under the pressure of the CRISPR-Cas system in the cellular population remain similar, depends solely on the nature of CRISPR-Cas and plasmid and does not depend on the initial plasmid distribution.

- The fraction of cells that retain plasmids, on the other hand, is affected by the initial distribution of plasmids upon activation of the CRISPR-Cas system.

- The distribution of plasmids under the pressure of CRISPR-Cas systems in the cellular population becomes bimodal, leading to the subpopulation with no plasmids and subpopulation that maintain plasmids

4.3.3 Follow-up experiments to check whether the stochastic kinetics of interference and replication explains plasmid survival

To better understand the composition of surviving colonies formed by CRISPR ON cells and to check the predictions of our model, in the follow–up experiments we replated the cells from both CRISPR ON and CRISPR OFF colonies onto three types of media (fig. 4.8): Plates with CRISPR-Cas inductor and antibiotics (orange), plates with only antibiotics (yellow), and plates with neither CRISPR-Cas inductor nor antibiotics (gray). In addition to counting colonies (CFU), we measured the number of plasmids in a colony using qPCR with plasmid-specific primers, normalized by the bacterial gyrA gene.

A qPCR analysis of colonies formed by CRISPR OFF transformants revealed that on average there are $[Pl]_{st} = 240 \pm 65$ plasmids per cell, which is consistent with published copy number values $[Pl]_{st}$ for pUC plasmids on which pG8 is based [162]. For the second plasmid pRSFG8, this number was 80 ± 25 which is also consistent with the published data [165]. In contrast, in colonies formed by the CRISPR-ON cells, an average cell had 0.13 ± 0.09 of the first plasmid and 0.44 ± 0.15 of the second one. (fig. 4.9). Superficially, the second observation contradicts the common sense (there should be at least a plasmid per cell to overcome antibiotic) and the nature of our model, since a low average plasmid copy number would make cells extremely vulnerable to a complete plasmid loss, making survival of a colony highly unlikely. Indeed, according to our predictions, the typical steady state number of plasmids in CRISPR ON cells $[Pl]_{eq}$ should be closer to that in CRISPR OFF cell ($[Pl]_{st}$) than to 1 (see fig. 4.7). This contradiction could be resolved assuming that cells in a CRISPR ON colony are heterogeneous, with some cells fully devoid of plasmid, while others maintaining the sufficient number of plasmids. To verify this, cells from both CRISPR ON and CRISPR

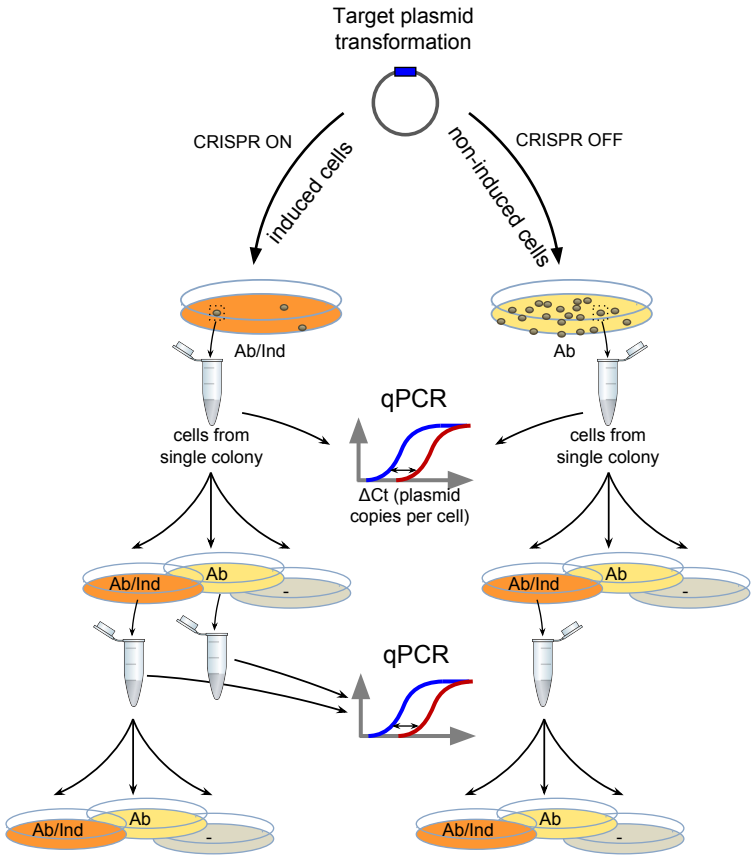

FIGURE 4.8: **Follow-up replating experiments scheme** In the follow-up experiments, both CRISPR ON and CRISPR OFF colonies were replated on three types of plates (second row of plates): with antibiotic and inducers (Ab/Ind, orange), antibiotic (Ab, yellow), and just the growth medium (grey) as a control of the total number of viable cells. To observe the subsequent dynamics of the induced cells (Ab/Ind), they were replated second time (third row of plates). To compare plasmid populations in colonies from Ab/Ind and Ab plates, the real-time PCR assay of random individual colonies was performed.

OFF transformed colonies were replated on antibiotic-free plates (to determine the total number of cells) and on plates with antibiotic (to determine the number of plasmid-bearing cells), the second line in fig. 4.8. For CRISPR OFF transformants, the number of colonies formed on plates with and without antibiotic was the same for both types of plasmids (second line in fig. 4.9, indicating that plasmids are stably maintained over the time of experiment even in the absence of selection. In contrast, only one out of a few thousand cells from CRISPR ON colonies grew on antibiotic-containing plates. This indicates that most cells in CRISPR ON colonies do not bear a plasmid and apparently survive due to an altruistic action of cells with plasmids. Thus, the model appears consistent with this replating experiment: Indeed, as we predict, the majority of transformed cell grown under the CRISPR ON conditions lose their plasmids, and the experimentally-derived average plasmid copy number in CRISPR ON survivors is large and similar to $[Pl]_{eq}$.

Another experiment aims to check our prediction that the number of CRISPR ON cells that lose plasmids strongly depends on the initial number of plasmids per cell. It consists of replating cells from CRISPR OFF colonies, which, as we know by now, have approximately $[Pl]_{st}$ plasmids each, on all three media and counting the colonies (fig. 4.8, right column). The results, shown in the second line of fig. 4.9 for both types of plasmids, indicate that turning CRISPR on does not decrease the number of colonies. Thus, the experiment confirms another model prediction that less than half of the cells with an initially large number of plasmids lose all plasmids in the long run. Evidently, such low-probable plasmid loss does affect colony formation. It is important to note, that, in general, the plasmids loss in the fraction of cells only indirectly affects the number of colonies observed. The fraction of colonies that can be observed on antibiotic i.e. the colonies that have some fraction of cells maintaining the plasmids will be significantly higher than the fraction of individual cells that holds the plasmid (see S1 appendix). As the colony

grows in cell number the probability that all on the cells will lose plasmids start approaching zero.

Finally, we set up a verification, albeit indirect, that the nature of plasmid distribution in plasmid-bearing CRISPR ON cells becomes the same after many generations for any initial number of plasmids. This correlates with the computational analysis showing that distribution of the plasmids in the cells holding the plasmids converge to the same distribution regardless of the initial plasmid distribution (see 4.7B,C). The cells initially grown either on CRISPR OFF or CRISPR ON medium and then replated on CRISPR ON medium are then replated the second time on all three media (fig. 4.8, right column). The bottom line in fig. 4.9 shows that the colony-forming capability of such twice replated cells is the same as that of the once replated CRISPR ON survivors. Furthermore, considering the differences between the number of colonies on CRISPR inductor and antibiotic medium, antibiotic medium, and just the growth medium as a proxy for plasmid distribution in cells, we conclude that this distribution is the same in CRISPR OFF $\rightarrow$ CRISPR ON cells and CRISPR ON $\rightarrow$ CRISPR ON cells. A small excess in CFU in the antibiotic-only medium compared to the CRISPR inductor plus antibiotic medium (visible in the third row in fig. 4.9 as the orange block is lower than the yellow one) can be attributed to a non-vanishing probability to loose all plasmids during each cell cycle, including the first ones that are crucial for colony formation. On contrary, such excess is not observed in the cells replated from CRISPR OFF colonies (three equal-height blocks in the second row in fig. 4.9, where such probability is very small due to a uniformly high ($\approx [Pl]_{st}$) an initial number of plasmids per cell.

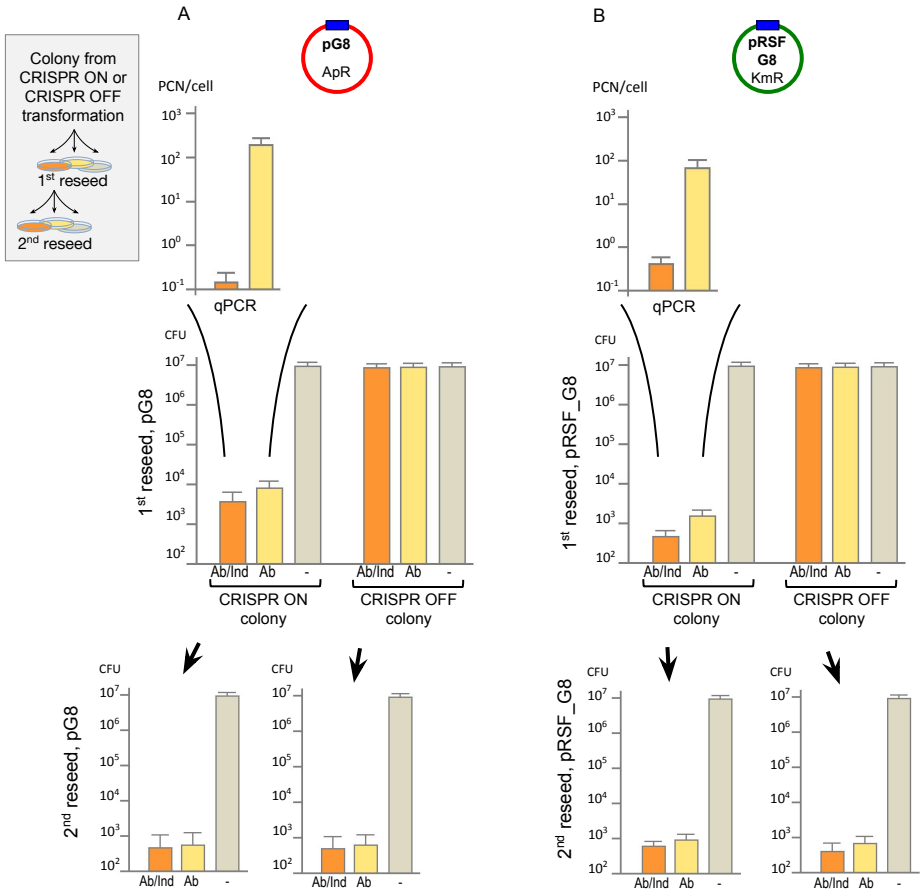

FIGURE 4.9: **Follow-up replating of CRISPR ON colonies results**. The transformants from CRISPR ON and CRISPR OFF cultures were used for the first replating on 3 types of plates (see fig. 4.8): Plates supplemented with antibiotics and inducers to maintain the CRISPR-Cas activity (orange), plates with antibiotic to determine the number of plasmid-bearing cells (yellow), and plates with growth medium only to estimate the total number of cells per colony (gray). To estimate the number of plasmids in the colonies the real-time PCR test was performed. The real-time PCR test shows the similar results for the transformants and cells from the colonies after replating. Column A and B describe sets of parallel experiments performed for pG8 and pRSFG8 plasmids.

4.4 Discussion

4.4.1 Summary of results

In this work, we reported the observation of survival of plasmids in a small fraction of cells, targeted by a perfectly functional CRISPR-Cas system, and provided quantitative explanations for this observation. Specifically:

- Performing targeted experiments, we ruled out several plausible reasons for CRISPR-Cas malfunction, such as mutations in spacers, protospacers, and *cas* genes.

- Instead, we explain the survival of plasmids as resulting from a rare sequence of stochastic plasmid duplication and CRISPR interference events. While the former should be on average less frequent than the latter, there is a finite probability that for a while the duplication exceeds cutting. Such duplication-dominant sequence of events can bring the plasmid population to the threshold, above which the average duplication rate is higher than the interference one. Above this threshold, the plasmid population becomes significantly more stable and harder to extinguish. Thus the majority of cells lose their plasmids failing to reach this threshold, while a few cells maintain on average the above-threshold plasmid population.

- A quantitative probabilistic model showed the viability of such explanation for plasmid survival. In addition to explaining the original experimental results, the model made testable predictions: The dependence of survival of plasmids on the initial number of plasmids in a cell, and the universality of distribution of plasmids per cell, which establishes itself after several generations independent of the initial number of plasmids

- These predictions have been checked and confirmed by specifically set follow-up experiments. We showed that cells with the initially large number of plasmids had a much higher probability to retain plasmids when CRISPR-Cas system was turned on. Yet the colony-forming capacity in the second and subsequent re-plating becomes the same and independent of the initial number of plasmids per cell, confirming the universality of distributions of plasmids

4.4.2 Kinetics of plasmid duplication and interference

Our model is based on rather simple assumptions about the plasmid kinetics, the Michaelis-Menten approximation and the logistic dynamics for the interference and duplication rates. Obviously, the modeling fidelity can be improved by utilizing experimentally-derived dependences of rates vs. plasmid copy number, or, at least, fitting the Michaelis-Menten and logistic constants to the experimental data. However, we believe that the salient features of the plasmid kinetics on which our explanation is based, the saturation of interference rate to a constant and the cessation of plasmid duplication for large plasmid copy number, illustrated in fig. 4.7A, would not change.

Since the stochastic survival of plasmids goes "against the odds" dictated by an excess of the interference rate over the duplication rate, the fraction of cells that retain plasmids falls dramatically with an increase of the rate-reversal threshold $[Pl]_{bif}$, (see fig. 4.7A). So a fairly delicate balance between the Michaelis-Menten and plasmid duplication rate constants, which determine $[Pl]_{bif}$, is required to observe the reported plasmid survival in a small fraction of cells. Obviously, at a single-cell level, the outcome of the plasmid-CRISPR conflict is purely random and all that can be predicted for a given cell is its probability to lose all or retain a certain number of plasmids.

4.4.3 Defense from viruses, horizontal gene transfer, and other evolutionary aspects

A natural question arises whether the same conclusions apply to CRISPR-Cas interaction with viruses. It is hard to give a definitive answer to this question as the viral replication apparently follows quite different kinetics than that of plasmids and, most importantly, it does not have to slow down when the number of viral copies reaches a certain threshold. Furthermore, a single escape mutation in viral protospacer would quickly propagate through the cell culture and save the viral population from extermination, making the detection of more delicate effects described in this work fairly complex. Finally, an infected cell usually receives just one copy of a virus while a cell inherits on average half of plasmids from its mother. Hence, if the probability for a virus to survive and replicate in a cell (quite low in the case of a single initial plasmid) multiplied by the size of a viral burst is smaller than one, the infection would not propagate and we would simply register an apparent defeat of viruses by CRISPR.

Nevertheless, an "imperfect" function of CRISPR-Cas system resulting in an incomplete extermination of viruses and plasmids can be a consequence of the same evolutionary principle: Increasing the rate of interference costs the cell not only some extra energy needed to produce additional copies of Cas proteins. A higher concentration of such complexes and their stronger binding to DNA carries inevitable risks of increasing chances of autoimmunity via binding to and attacking self-DNA. Thus the evolutionary optimization of CRISPR-Cas interference rate would probably not go beyond some intermediate protection level, giving only a majority of cells rather than everyone, means to eliminate foreign mobile genetic elements. Thus, in a fraction of cells plasmids and viruses could survive, either by the stochastic mechanism proposed above or due to a cell to cell (or single-cell temporal) [166]

variability in concentrations of Cas proteins and thus interference rates. Besides, such "imperfectly" functioning CRISPR-Cas system could allow plasmids that carry benefits for cells to take a hold and then proliferate in the population, thus lowering the efficiency but not entirely blocking HGT.

4.4.4 Comparison with the previous observations

Our experimental results are seemingly controversial to the current typical view on the interaction between plasmids and CRISPR-Cas system. It was considered that the only way the plasmid could escape the interference is through mutation of one of the components of this interaction, either spacer, spacer or mutation in the CRISPR-Cas system itself [106]. Here we propose that this plasmid escape could be reached for kinetic reasons. However, this does not abolish previous observations, which, presumably, correspond to the CRISPR-plasmid interaction with different kinetic properties. Indeed, if the CRISPR-Cas interference rate is high compared to plasmid replication rate there is no "plasmid stability window" (see Fig. 4.6A) and plasmids are to be eliminated unless there is something broken in the CRISPR-Cas interference. Moreover, these mechanisms of plasmid escape could be complementary. Plasmid stability window could become a beachhead for further penetration of plasmid through mutation and escape of CRISPR-Cas system. It allows the plasmid to survive for high enough number of replication to allow mutation. On the other hand, the most recent results in *Enterococcus faecalis* by [167] shows that indeed the native level of expression is both non-lethal in targeting self DNA and potentially could not inhibit horizontal gene transfer. At the same time, the over-expression of /textitcas genes leads to the elimination of such an effect. In the framework of our model, we propose that it could be due to the shift in the kinetic curve of interference and following change in the interaction between the plasmids and CRISPR-Cas.

While in the native system the interference rate is low and the system is in "plasmid wins" regime in the over-expressed system the interference rate is high and the system is in the "CRISPR wins" regime.

4.4.5 New view on the interaction between CRISPR and CRISPR targets

In most research works CRISPR-Cas targets are treated as passive elements. Partially it might be driven by the studies of the overexpressed CRISPR-Cas models when the interference rate is so high that the elimination of the target is almost inevitable. However, in the native systems the Cas protein expression rate is significantly lower thus the competition between the CRISPR-Cas system and its target can start playing the crucial role. We propose that the CRISPR target should not be viewed as a passive element, only waiting to be destroyed, but is an active system that could compete with and out-compete CRISPR-Cas interference. This paradigm leads to the hypothesis that the nature of the CRISPR target and its replication kinetics should affect the outcome of the CRISPR-target competition and different targets can compete differently with the same CRISPR-Cas system.

4.4.6 Processes potentially affecting the outcome that are not simulated

While the model manages to capture the main features of the colonies behavior we found challenging to replicate the quantitatively exact scenario as in the experiments. We contribute it to several features that were not modeled yet shapes the colony behavior and cellular growth. The first is the spatial structure of the bacterial colonies. It is well known that bacterial colonies and biofilms are not homogeneous and form micro-environments that have

complex spatial structure [168, 169] that may also include cooperative antibiotic resistance [170]. Also, there were studies that show that spatial structure drastically affects the behavior of the CRISPR-Cas systems and contribute to the CRISPR-Cas system evolution [78, 129]. The second is the effect of the plasmids presence on cellular growth. It has been shown in a series of different experiments that holding a high-copy-number plasmid might affect the cellular growth. While it remains unclear whether there is a divergence of cellular replication rate based on the plasmid copy number in case of our experiments this could, in theory, contribute to the bias in the experimental outcome.

Chapter 5

Discussions and Conclusion

In this work, we analyzed two related topics through mathematical modeling. One was the CRISPR array composition and in particular the number of spacers in the CRISPR array that maximizes the protection against the foreign genetic elements. The second is the condition when foreign genetic elements can overcome the CRISPR interference despite the activity of the CRISPR-Cas system. Together they advance the theoretical understanding of the CRISPR-Cas protection and factors that toggle the balance of the competition between CRISPR-Cas and foreign genetic elements to one side or another.

Analyzing the optimal spacer array composition, we came to several important conclusions. First and foremost, we found that for a constant environmental condition there is a non-trivial (i.e. non-zero and non-infinite) optimum in the number of spacers that maximizes the host cell protection. It is dictated by various factors such as the efficiency of CRISPR-Cas recognition and nuclease machinery, abundance of crRNA with spacers located in different positions of the array, viral mutation rate, viral abundance etc. Surprisingly, higher efficiencies of the CRISPR-Cas system lead to longer arrays while the higher viral mutation rate makes shorter arrays optimal from the point of efficiency. Yet generally, short arrays have shown to be in general inefficient against viral attacks, and such arrays should rely on either constant uptake of new spacers or more sophisticated mechanisms of spacer update such as primed adaptation and constant purging of older arrays. Also, it has

been shown that proportionally higher production of crRNA from the newer spacers is beneficial for CRISPR-Cas system, leading to both higher efficiency of CRISPR-Cas system and higher CRISPR-Cas efficiency robustness against the changes in the number of spacers in the CRISPR array. This suggests that palindromic nature of the CRISPR repeats might not only play a structural role but also have a regulatory function in Type I and Type III systems (by terminating transcription) and could be subject to evolutionary pressure in order to adapt to changing viral environment.

In general, this model provides a different view on the spacer array composition problem. Instead of the focus on the dynamics of the uptake and loss of the spacers typically analyzed in other works, our approach allowed us to perform a more detailed analysis of the CRISPR-Cas efficiency under given conditions.

Analyzing the interaction between foreign genetic elements (plasmids in particular) and CRISPR-Cas systems, we focused more on the kinetics and stochastic nature of the process of the competition between plasmid replication and CRISPR interference. We have experimentally shown that despite the common view that plasmids can escape the CRISPR-Cas immunity only through mutations in targeted protospacers and their PAMs, plasmids can avoid CRISPR interference via other mechanisms. We have demonstrated that ongoing plasmid replication and CRISPR-Cas interference could lead to various scenarios, including one of plasmid overcoming the CRISPR-Cas immunity through rapid replication. However, even if the plasmid population seems doomed as the interference dominates, for a low number of plasmids, a "plasmid stability window" – the intermediate range of plasmid copy number where plasmid replication rate is higher than CRISPR-Cas interference – exists. Such condition leads to unique bimodality in the cellular population with one subpopulation losing all plasmids while the other maintaining them.

In general, we propose that the interaction between the CRISPR-Cas system and foreign genetic element, as any birth-death process with a few players, should be viewed as a stochastic process. While most of the modern works treat foreign genetic elements as passive targets of CRISPR-Cas, in reality the situation is certainly more complex. The nature of replication and behavior of the particular genetic element affect the outcome of the competition between it and CRISPR-Cas and could lead to complex non-deterministic outcomes such as bimodality in the population.

Bibliography

[1] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. a. Romero, and P. Horvath, "CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes," *Science*, vol. 315, pp. 1709–1712, mar 2007.

[2] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and E. Soria, "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements," *Journal of Molecular Evolution*, vol. 60, no. 2, pp. 174–182, 2005.

[3] 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, J. van der Oost, and E. V. Koonin, "Evolution and classification of the CRISPR–Cas systems," *Nature Reviews Microbiology*, vol. 9, pp. 467–477, jun 2011.

[4] 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, P. Horvath, S. Moineau, F. J. M. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. van der Oost, R. Backofen, and E. V. Koonin, "An updated evolutionary classification of CRISPR–Cas systems," *Nature Reviews Microbiology*, vol. 13, pp. 722–736, sep 2015.

[5] E. V. Koonin, K. S. Makarova, and F. Zhang, "Diversity, classification and evolution of CRISPR-Cas systems," *Current Opinion in Microbiology*, vol. 37, pp. 67–78, jun 2017.

[6] I. Yosef, M. G. Goren, and U. Qimron, "Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli," *Nucleic Acids Research*, vol. 40, pp. 5569–5576, jul 2012.

[7] Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, and A. Nakata, "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.," *Journal of Bacteriology*, vol. 169, no. 12, pp. 5429–5433, 1987.

[8] A. Bolotin, B. Quinquis, A. Sorokin, and S. Dusko Ehrlich, "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin," *Microbiology*, vol. 151, no. 8, pp. 2551–2561, 2005.

[9] C. Pourcel, G. Salvignol, and G. Vergnaud, "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies," *Microbiology*, vol. 151, no. 3, pp. 653–663, 2005.

[10] S. Shmakov, O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, and E. V. Koonin, "Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems," *Molecular Cell*, vol. 60, no. 3, pp. 385–397, 2015.

[11] R. N. Jackson and B. Wiedenheft, "A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses," *Molecular Cell*, vol. 58, pp. 722–728, jun 2015.

[12] K. S. Makarova, L. Aravind, Y. I. Wolf, and E. V. Koonin, "Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR- Cas systems," *Biology Direct*, vol. 6, no. 1, p. 38, 2011.

[13] T. Sinkunas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys, "Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system," *The EMBO Journal*, vol. 30, pp. 1335–1342, apr 2011.

[14] K. Chylinski, K. S. Makarova, E. Charpentier, and E. V. Koonin, "Classification and evolution of type II CRISPR-Cas systems," *Nucleic Acids Research*, vol. 42, pp. 6091–6105, jun 2014.

[15] V. Anantharaman, K. S. Makarova, A. M. Burroughs, E. V. Koonin, and L. Aravind, "Comprehensive analysis of the HEPN superfamily: Identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing," *Biology Direct*, vol. 8, no. 1, p. 1, 2013.

[16] J. Carte, R. Wang, H. Li, R. M. Terns, and M. P. Terns, "Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes," *Genes and Development*, vol. 22, no. 24, pp. 3489–3496, 2008.

[17] 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, "Small CRISPR RNAs guide antiviral defense in prokaryotes.," *Science (New York, N.Y.)*, vol. 321, pp. 960–4, aug 2008.

[18] L. A. Marraffini, "CRISPR-Cas immunity in prokaryotes," *Nature*, vol. 526, pp. 55–61, oct 2015.

[19] R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, and J. A. Doudna, "Sequence- and structure-specific RNA processing by a CRISPR endonuclease," *Science*, vol. 329, no. 5997, pp. 1355–1358, 2010.

[20] E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel, and E. Charpentier, "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III," *Nature*, vol. 471, no. 7340, pp. 602–607, 2011.

[21] A. Hatoum-Aslan, I. Maniv, and L. A. Marraffini, "Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site," *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 21218–21222, 2011.

[22] R. D. Sokolowski, S. Graham, and M. F. White, "Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system," *Nucleic Acids Research*, vol. 42, no. 10, pp. 6532–6541, 2014.

[23] 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, "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*, vol. 108, no. 25, pp. 10098–10103, 2011.

[24] D. G. Sashital, B. Wiedenheft, and J. A. Doudna, "Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System," *Molecular Cell*, vol. 46, no. 5, pp. 606–615, 2012.

[25] M. L. Hochstrasser, D. W. Taylor, P. Bhat, C. K. Guegler, S. H. Sternberg, E. Nogales, and J. A. Doudna, "CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference," *Proceedings of the National Academy of Sciences*, vol. 111, no. 18, pp. 6618–6623, 2014.

[26] D. C. Swarts, J. van der Oost, and M. Jinek, "Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a," *Molecular Cell*, vol. 66, pp. 221–233.e4, apr 2017.

[27] H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau, "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus," *Journal of Bacteriology*, vol. 190, no. 4, pp. 1390–1400, 2008.

[28] R. Sapranauskas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys, "The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9275–9282, 2011.

[29] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek, "Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease," *Nature*, vol. 513, no. 7519, pp. 569–573, 2014.

[30] J. E. Garneau, M. È. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadán, and S. Moineau, "The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA," *Nature*, vol. 468, no. 7320, pp. 67–71, 2010.

[31] H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, and O. Nureki, "Crystal structure of Cas9 in complex with guide RNA and target DNA," *Cell*, vol. 156, no. 5, pp. 935–949, 2014.

[32] L. A. Marraffini and E. J. Sontheimer, "Self versus non-self discrimination during CRISPR RNA-directed immunity," *Nature*, vol. 463, no. 7280, pp. 568–571, 2010.

[33] P. Samai, N. Pyenson, W. Jiang, G. W. Goldberg, A. Hatoum-Aslan, and L. A. Marraffini, "Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity," *Cell*, vol. 161, no. 5, pp. 1164–1174, 2015.

[34] H. Nishimasu and O. Nureki, "Structures and mechanisms of CRISPR RNA-guided effector nucleases," *Current Opinion in Structural Biology*, vol. 43, pp. 68–78, 2017.

[35] R. H. J. Staals, Y. Agari, S. Maki-Yonekura, Y. Zhu, D. W. Taylor, E. Van-Duijn, A. Barendregt, M. Vlot, J. J. Koehorst, K. Sakamoto, A. Masuda, N. Dohmae, P. Schaap, J. A. Doudna, A. J. R. Heck, K. Yonekura, J. Van der Oost, and A. Shinkai, "Structure and Activity of the RNA-Targeting Type III-B CRISPR-Cas Complex of Thermus thermophilus," *Molecular Cell*, vol. 52, no. 1, pp. 135–145, 2013.

[36] J. R. Elmore, N. F. Sheppard, N. Ramia, T. Deighan, H. Li, R. M. Terns, and M. P. Terns, "Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR–Cas system," *Genes and Development*, vol. 30, no. 4, pp. 447–459, 2016.

[37] Ü. Pul, R. Wurm, Z. Arslan, R. Geißen, N. Hofmann, and R. Wagner, "Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS," *Molecular Microbiology*, vol. 75, no. 6, pp. 1495–1512, 2010.

[38] E. R. Westra, Ü. Pul, N. Heidrich, M. M. Jore, M. Lundgren, T. Stratmann, R. Wurm, A. Raine, M. Mescher, L. Van Heereveld, M. Mastop, E. G. H. Wagner, K. Schnetz, J. Van Der Oost, R. Wagner, and S. J. J. Brouns, "H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO," *Molecular Microbiology*, vol. 77, no. 6, pp. 1380–1393, 2010.

[39] T. Liu, Y. Li, X. Wang, Q. Ye, H. Li, Y. Liang, Q. She, and N. Peng, "Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition," *Nucleic Acids Research*, vol. 43, no. 2, pp. 1044–1055, 2015.

[40] T. Stratmann, Ü. Pul, R. Wurm, R. Wagner, and K. Schnetz, "RcsB-BglJ activates the Escherichia coli leuO gene, encoding an H-NS antagonist and pleiotropic regulator of virulence determinants," *Molecular Microbiology*, vol. 83, no. 6, pp. 1109–1123, 2012.

[41] O. O. Abudayyeh, J. S. Gootenberg, S. Konermann, J. Joung, I. M. Slaymaker, D. B. T. Cox, S. Shmakov, K. S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E. S. Lander, E. V. Koonin, and F. Zhang, "C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector," *Science*, vol. 353, p. aaf5573, aug 2016.

[42] G. Amitai and R. Sorek, "CRISPR–Cas adaptation: insights into the mechanism of action," *Nature Reviews Microbiology*, vol. 14, pp. 67–76, jan 2016.

[43] A. Levy, M. G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, and R. Sorek, "CRISPR adaptation biases explain preference for acquisition of foreign DNA," *Nature*, vol. 520, pp. 505–510, apr 2015.

[44] M. R. Singleton, M. S. Dillingham, M. Gaudier, S. C. Kowalczykowski, and D. B. Wigley, "Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks," *Nature*, vol. 432, no. 7014, pp. 187–193, 2004.

[45] P. E. Boehmer and P. T. Emmerson, "The RecB subunit of the Escherichia coli RecBCD enzyme couples ATP hydrolysis to DNA unwinding," *Journal of Biological Chemistry*, vol. 267, no. 7, pp. 4981–4987, 1992.

[46] R. Benzinger, L. W. Enquist, and A. Skalka, "Transfection of Escherichia coli spheroplasts. V. Activity of recBC nuclease in rec+ and rec minus spheroplasts measured with different forms of bacteriophage DNA.," *Journal of virology*, vol. 15, pp. 861–71, apr 1975.

[47] D. Lackey and S. Linn, "[4] Assay for type II restriction endonucleases using the Escherichia coli rec BC DNase and duplex circular DNA," in *Methods in Enzymology*, vol. 65, pp. 26–28, 1980.

[48] A. Miranda and A. Kuzminov, "Chromosomal lesion suppression and removal in Escherichia coli via linear DNA degradation.," *Genetics*, vol. 163, pp. 1255–71, apr 2003.

[49] B. Michel, H. Boubakri, Z. Baharoglu, M. LeMasson, and R. Lestini, "Recombination proteins and rescue of arrested replication forks.," *DNA repair*, vol. 6, pp. 967–80, jul 2007.

[50] M. S. Dillingham and S. C. Kowalczykowski, "RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks," *Microbiology and Molecular Biology Reviews*, vol. 72, pp. 642–671, dec 2008.

[51] G. R. Smith, "How RecBCD Enzyme and Chi Promote DNA Break Repair and Recombination: a Molecular Biologist's View," *Microbiology and Molecular Biology Reviews*, vol. 76, pp. 217–228, jun 2012.

[52] M. El Karoui, V. Biaudet, S. Schbath, and a. Gruss, "Characteristics of Chi distribution on different bacterial genomes.," *Research in microbiology*, vol. 150, no. 9-10, pp. 579–87, 1999.

[53] S. N. Kieper, C. Almendros, J. Behler, R. E. McKenzie, F. L. Nobrega, A. C. Haagsma, J. N. Vink, W. R. Hess, and S. J. Brouns, "Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation," *Cell Reports*, vol. 22, no. 13, pp. 3377–3384, 2018.

[54] J. K. Nuñez, A. S. Y. Lee, A. Engelman, and J. A. Doudna, "Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity," *Nature*, vol. 519, pp. 193–198, mar 2015.

[55] J. Wang, J. Li, H. Zhao, G. Sheng, M. Wang, M. Yin, and Y. Wang, "Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems," *Cell*, vol. 163, no. 4, pp. 840–853, 2015.

[56] Z. Arslan, V. Hermanns, R. Wurm, R. Wagner, and Ü. Pul, "Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system," *Nucleic Acids Research*, vol. 42, pp. 7884–7893, jul 2014.

[57] R. Heler, P. Samai, J. W. Modell, C. Weiner, G. W. Goldberg, D. Bikard, and L. A. Marraffini, "Cas9 specifies functional viral targets during CRISPR – Cas adaptation," *Nature*, 2015.

[58] K. A. Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, and E. Semenova, "Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system," *Nature Communications*, vol. 3, no. May, pp. 945–947, 2012.

[59] A. Krivoy, M. Rutkauskas, K. Kuznedelov, O. Musharova, C. Rouillon, K. Severinov, and R. Seidel, "Primed CRISPR adaptation in Escherichia coli cells does not depend on conformational changes in the Cascade effector complex detected in Vitro," *Nucleic Acids Research*, pp. 1–12, mar 2018.

[60] K. Severinov, I. Ispolatov, and E. Semenova, "The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense," *Frontiers in Molecular Biosciences*, vol. 3, p. 45, aug 2016.

[61] E. Semenova, E. Savitskaya, O. Musharova, A. Strotskaya, D. Vorontsova, K. A. Datsenko, M. D. Logacheva, and K. Severinov, "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*, vol. 113, no. 27, pp. 7626–7631, 2016.

[62] T. Künne, S. N. Kieper, J. W. Bannenberg, A. I. Vogel, W. R. Miellet, M. Klein, M. Depken, M. Suarez-Diez, and S. J. Brouns, "Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation," *Molecular Cell*, vol. 63, no. 5, pp. 852–864, 2016.

[63] C. Xue, N. R. Whitis, and D. G. Sashital, "Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity," *Molecular Cell*, vol. 64, pp. 826–834, nov 2016.

[64] S. Redding, S. H. Sternberg, B. Wiedenheft, A. Jennifer, E. C. Greene, S. Redding, S. H. Sternberg, M. Marshall, B. Gibb, P. Bhat, and C. K. Guegler, "Supplemental Information: Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System," *Cell*, vol. 163, no. 4, pp. 1–12, 2015.

[65] B. Koskella and M. A. Brockhurst, "Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities," *FEMS Microbiology Reviews*, vol. 38, no. 5, pp. 916–931, 2014.

[66] E. V. Koonin and Y. I. Wolf, "Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus–host coevolution," *Mol. BioSyst.*, vol. 11, no. 1, pp. 20–27, 2015.

[67] L. van Valen, "A new evolutionary law," *Evolutionary Theory*, vol. 1, pp. 1–30, apr 1973.

[68] E. Savitskaya, E. Semenova, V. Dedkov, A. Metlitskaya, and K. Severinov, "High-throughput analysis of type I-E CRISPR / Cas spacer acquisition in E . coli," *RNA biology*, vol. 10, no. 5, pp. 716–725, 2013.

[69] A. D. Weinberger, C. L. Sun, M. M. Pluciński, V. J. Denef, B. C. Thomas, P. Horvath, R. Barrangou, M. S. Gilmore, W. M. Getz, and J. F. Banfield, "Persisting viral sequences shape microbial CRISPR-based immunity," *PLoS Computational Biology*, vol. 8, no. 4, 2012.

[70] P. Horvath, D. A. Romero, A. C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou, "Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus," *Journal of Bacteriology*, vol. 190, no. 4, pp. 1401–1412, 2008.

[71] M. J. Lopez-Sanchez, E. Sauvage, V. Da Cunha, D. Clermont, E. Ratsima Hariniaina, B. Gonzalez-Zorn, C. Poyart, I. Rosinski-Chupin, and P. Glaser, "The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome," *Molecular Microbiology*, vol. 85, no. 6, pp. 1057–1071, 2012.

[72] C. L. Sun, B. C. Thomas, R. Barrangou, and J. F. Banfield, "Metagenomic reconstructions of bacterial CRISPR loci constrain population histories," *ISME Journal*, vol. 10, no. 4, pp. 858–870, 2016.

[73] E. Laanto, V. Hoikkala, J. Ravantti, and L. R. Sundberg, "Long-term genomic coevolution of host-parasite interaction in the natural environment," *Nature Communications*, vol. 8, no. 1, 2017.

[74] A. Strotskaya, E. Savitskaya, A. Metlitskaya, N. Morozova, K. A. Datsenko, E. Semenova, and K. Severinov, "The action of Escherichia coli CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies," *Nucleic acids research*, vol. 45, pp. 1946–1957, jan 2017.

[75] S. Luria and M. Delbrück, "Mutations of Bacteria from Virus Sensitivity to Virus Resistance.," *Genetics*, vol. 28, no. 6, pp. 491–511, 1943.

[76] J. Lederberg and E. M. Lederberg, "Replica Plating and Indirect Selection of Bacterial Mutants," *Journal of Bacteriology*, vol. 63, no. 3, pp. 399–406, 1952.

[77] E. V. Koonin and Y. I. Wolf, "Is evolution Darwinian or/and Lamarckian?," *Biology Direct*, vol. 4, no. 1, p. 42, 2009.

[78] J. O. Haerter and K. Sneppen, "Spatial Structure and Lamarckian Adaptation Explain Extreme Genetic Diversity at CRISPR Locus," *mBio*, vol. 3, pp. e00126–12–e00126–12, jul 2012.

[79] E. V. Koonin and Y. I. Wolf, "Just how Lamarckian is CRISPR-Cas immunity: the continuum of evolvability mechanisms," *Biology Direct*, vol. 11, no. 1, p. 9, 2016.

[80] L. M. Childs, N. L. Held, M. J. Young, R. J. Whitaker, and J. S. Weitz, "Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin," *Evolution*, vol. 66, pp. 2015–2029, jul 2012.

[81] A. P. Hynes, M. Villion, and S. Moineau, "Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages," *Nature Communications*, vol. 5, no. May, pp. 1–6, 2014.

[82] K. L. Maxwell, "The Anti-CRISPR Story: A Battle for Survival," *Molecular Cell*, vol. 68, no. 1, pp. 8–14, 2017.

[83] J. Wang, J. Ma, Z. Cheng, X. Meng, L. You, M. Wang, X. Zhang, and Y. Wang, "A CRISPR evolutionary arms race: Structural insights into viral anti-CRISPR/Cas responses," *Cell Research*, vol. 26, no. 10, pp. 1165–1168, 2016.

[84] J. Bondy-Denomy, A. Pawluk, K. L. Maxwell, and A. R. Davidson, "Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system," *Nature*, vol. 493, pp. 429–432, dec 2013.

[85] A. Pawluk, J. Bondy-Denomy, V. H. W. Cheung, K. L. Maxwell, and A. R. Davidson, "A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa," *mBio*, vol. 5, pp. e00896–14–e00896–14, apr 2014.

[86] J. Bondy-Denomy, B. Garcia, S. Strum, M. Du, M. F. Rollins, Y. Hidalgo-Reyes, B. Wiedenheft, K. L. Maxwell, and A. R. Davidson, "Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins," *Nature*, vol. 526, no. 7571, pp. 136–139, 2015.

[87] D. Dong, M. Guo, S. Wang, Y. Zhu, S. Wang, Z. Xiong, J. Yang, Z. Xu, and Z. Huang, "Structural basis of CRISPR-SpyCas9 inhibition by an anti-CRISPR protein," *Nature*, vol. 546, no. 7658, pp. 436–439, 2017.

[88] L. B. Harrington, K. W. Doxzen, E. Ma, J. J. Liu, G. J. Knott, A. Edraki, B. Garcia, N. Amrani, J. S. Chen, J. C. Cofsky, P. J. Kranzusch, E. J. Sontheimer, A. R. Davidson, K. L. Maxwell, and J. A. Doudna, "A Broad-Spectrum Inhibitor of CRISPR-Cas9," *Cell*, vol. 170, no. 6, pp. 1224–1233.e15, 2017.

[89] X. Wang, D. Yao, J. G. Xu, A. R. Li, J. Xu, P. Fu, Y. Zhou, and Y. Zhu, "Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3," *Nature Structural and Molecular Biology*, vol. 23, no. 9, pp. 868–870, 2016.

[90] M. S. Kumar, J. B. Plotkin, and S. Hannenhalli, "Regulated CRISPR Modules Exploit a Dual Defense Strategy of Restriction and Abortive Infection in a Model of Prokaryote-Phage Coevolution," *PLoS Computational Biology*, vol. 11, no. 11, pp. 1–25, 2015.

[91] S. J. Labrie, J. E. Samson, and S. Moineau, "Bacteriophage resistance mechanisms," 2010.

[92] M. C. Chopin, A. Chopin, and E. Bidnenko, "Phage abortive infection in lactococci: Variations on a theme," 2005.

[93] A. Stern, L. Keren, O. Wurtzel, G. Amitai, and R. Sorek, "Self-targeting by CRIPR: gene regulation or autoimmunity?," *Trends Genet.*, vol. 26, pp. 335–340, 2010.

[94] E. V. Koonin and Y. I. Wolf, "Evolution of microbes and viruses: a paradigm shift in evolutionary biology?," *Frontiers in Cellular and Infection Microbiology*, vol. 2, no. September, pp. 1–15, 2012.

[95] S. M. Soucy, J. Huang, and J. P. Gogarten, "Horizontal gene transfer: Building the web of life," *Nature Reviews Genetics*, vol. 16, no. 8, pp. 472–482, 2015.

[96] C. M. Thomas and K. M. Nielsen, "Mechanisms of, and barriers to, horizontal gene transfer between bacteria," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 711–721, 2005.

[97] E. L. Tatum and J. Lederberg, "Gene Recombination in the Bacterium Escherichia coli.," *Journal of bacteriology*, vol. 53, pp. 673–84, jun 1947.

[98] N. T. Perna, G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Pósfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner, "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7," *Nature*, vol. 409, pp. 529–533, jan 2001.

[99] T. J. Treangen and E. P. C. Rocha, "Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes," *PLoS Genetics*, vol. 7, p. e1001284, jan 2011.

[100] A. Norman, L. H. Hansen, and S. J. Sorensen, "Conjugative plasmids: vessels of the communal gene pool," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1527, pp. 2275–2289, 2009.

[101] C. Johnston, B. Martin, G. Fichant, P. Polard, and J. P. Claverys, "Bacterial transformation: Distribution, shared mechanisms and divergent control," *Nature Reviews Microbiology*, vol. 12, no. 3, pp. 181–196, 2014.

[102] A. M. Friedlander, "DNA release as a direct measure of microbial killing by phagocytes," *Infection and Immunity*, vol. 22, no. 1, pp. 148–154, 1978.

[103] L. A. Marraffini and E. J. Sontheimer, "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.," *Science*, vol. 322, no. 5909, pp. 1843–1845, 2008.

[104] O. Popa and T. Dagan, "Trends and barriers to lateral gene transfer in prokaryotes," *Current Opinion in Microbiology*, vol. 14, no. 5, pp. 615–623, 2011.

[105] D. Bikard, A. Hatoum-Aslan, D. Mucida, and L. A. Marraffini, "CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection," *Cell Host and Microbe*, vol. 12, no. 2, pp. 177–186, 2012.

[106] W. Jiang, I. Maniv, F. Arain, Y. Wang, B. R. Levin, and L. A. Marraffini, "Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids," *PLoS Genetics*, vol. 9, p. e1003844, sep 2013.

[107] Y. Zhang, N. Heidrich, B. J. Ampattu, C. W. Gunderson, H. S. Seifert, C. Schoen, J. Vogel, and E. J. Sontheimer, "Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis," *Molecular Cell*, vol. 50, pp. 488–503, may 2013.

[108] A. Brodt, M. N. Lurie-Weinberger, and U. Gophna, "CRISPR loci reveal networks of gene exchange in archaea," *Biol Direct*, vol. 6, no. 1, p. 65, 2011.

[109] U. Gophna, D. M. Kristensen, Y. I. Wolf, O. Popa, C. Drevet, and E. V. Koonin, "No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales.," *The ISME journal*, vol. 9, no. 9, pp. 2021–7, 2015.

[110] N. L. Held, A. Herrera, H. C. Quiroz, and R. J. Whitaker, "CRISPR associated diversity within a population of Sulfolobus islandicus," *PLoS ONE*, vol. 5, no. 9, 2010.

[111] K. R. Hargreaves, C. O. Flores, T. D. Lawley, and M. R. J. Clokie, "Abundant and Diverse Clustered Regularly Interspaced Short Palindromic Repeat Spacers in Clostridium difficile Strains and Prophages Target Multiple Phage Types within This Pathogen," *mBio*, vol. 5, pp. e01045–13–e01045–13, aug 2014.

[112] S. A. Shmakov, V. Sitnik, K. S. Makarova, Y. I. Wolf, K. V. Severinov, and E. V. Koonin, "The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes," *mBio*, vol. 8, pp. e01397–17, nov 2017.

[113] G. C. McGhee and G. W. Sundin, "Erwinia amylovora CRISPR elements provide new tools for evaluating strain diversity and for microbial source tracking," *PLoS ONE*, vol. 7, no. 7, 2012.

[114] 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, K. Miller, M. C. Enright, N. Mugnier, D. Brami, S. Schicklin, M. Felderman, A. S. Schwartz, T. H. Richardson, T. C. Peterson, B. Hubby, and K. C. Cady, "Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant Pseudomonas aeruginosa.," *mBio*, vol. 6, no. 6, pp. 1–13, 2015.

[115] E. R. Westra, A. Buckling, and P. C. Fineran, "CRISPR-Cas systems: beyond adaptive immunity.," *Nature reviews. Microbiology*, vol. 12, no. 5, pp. 317–26, 2014.

[116] A. A. Dominguez, W. A. Lim, and L. S. Qi, "Beyond editing: Repurposing CRISPR-Cas9 for precision genome regulation and interrogation," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 1, pp. 5–15, 2016.

[117] R. S. Pillai, S. N. Bhattacharyya, and W. Filipowicz, "Repression of protein synthesis by miRNAs: how many mechanisms?," 2007.

[118] M. E. Zegans, J. C. Wagner, K. C. Cady, D. M. Murphy, J. H. Hammond, and G. A. O'Toole, "Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa," *Journal of Bacteriology*, vol. 91, no. 1, pp. 210–219, 2009.

[119] C. R. Hale, P. Zhao, S. Olson, M. O. Duff, B. R. Graveley, L. Wells, R. M. Terns, and M. P. Terns, "RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex," *Cell*, vol. 139, no. 5, pp. 945–956, 2009.

[120] M. Babu, N. Beloglazova, R. Flick, C. Graham, T. Skarina, B. Nocek, A. Gagarinova, O. Pogoutse, G. Brown, A. Binkowski, S. Phanse, A. Joachimiak, E. V. Koonin, A. Savchenko, A. Emili, J. Greenblatt, A. M. Edwards, and A. F. Yakunin, "A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair," *Molecular Microbiology*, vol. 79, pp. 484–502, jan 2011.

[121] 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, "Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands," *PLoS Genetics*, vol. 9, no. 4, 2013.

[122] P. F. Vale and T. J. Little, "CRISPR-mediated phage resistance and the ghost of coevolution past," *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1691, pp. 2097–2103, 2010.

[123] P. F. Vale, G. Lafforgue, F. Gatchitch, R. Gardan, S. Moineau, and S. Gandon, "Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus," *Proceedings of the Royal Society B: Biological Sciences*, vol. 282, p. 20151270, aug 2015.

[124] C. H. Kuo and H. Ochman, "Deletional Bias across the Three Domains of Life," *Genome Biology and Evolution*, vol. 1, pp. 145–152, may 2010.

[125] D. Rath, L. Amlinger, A. Rath, and M. Lundgren, "The CRISPR-Cas immune system: Biology, mechanisms and applications," *Biochimie*, vol. 117, pp. 119–128, 2015.

[126] Y. Wei, R. M. Terns, and M. P. Terns, "Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation," *Genes & Development*, vol. 29, pp. 356–361, feb 2015.

[127] E. R. Westra, S. Van houte, S. Oyesiku-Blakemore, B. Makin, J. M. Broniewski, A. Best, J. Bondy-Denomy, A. Davidson, M. Boots, and A. Buckling, "Parasite exposure drives selective evolution of constitutive versus inducible defense," *Current Biology*, vol. 25, no. 8, pp. 1043–1049, 2015.

[128] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, "Why the Monte Carlo method is so important today," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 386–392, 2014.

[129] J. O. Haerter, A. Trusina, and K. Sneppen, "Targeted Bacterial Immunity Buffers Phage Diversity," *Journal of Virology*, vol. 85, pp. 10554–10560, oct 2011.

[130] L. M. Childs, W. E. England, M. J. Young, J. S. Weitz, and R. J. Whitaker, "CRISPR-Induced Distributed Immunity in Microbial Populations," *PLoS ONE*, vol. 9, p. e101710, jul 2014.

[131] J. He and M. W. Deem, "Heterogeneous Diversity of Spacers within CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)," *Physical Review Letters*, vol. 105, p. 128102, sep 2010.

[132] P. Han, L. R. Niestemski, J. E. Barrick, and M. W. Deem, "Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system," *Physical Biology*, vol. 10, p. 025004, mar 2013.

[133] B. R. Levin, "Nasty Viruses, Costly Plasmids, Population Dynamics, and the Conditions for Establishing and Maintaining CRISPR-Mediated Adaptive Immunity in Bacteria," *PLoS Genetics*, vol. 6, p. e1001171, oct 2010.

[134] B. R. Levin, S. Moineau, M. Bushman, and R. Barrangou, "The Population and Evolutionary Dynamics of Phage and Bacteria with CRISPR-Mediated Immunity," *PLoS Genetics*, vol. 9, no. 3, 2013.

[135] A. D. Weinberger, Y. I. Wolf, A. E. Lobkovsky, M. S. Gilmore, and E. V. Koonin, "Viral diversity threshold for adaptive immunity in prokaryotes.," *mBio*, vol. 3, no. 6, pp. 1–10, 2012.

[136] J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, "Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context," *Journal of Bacteriology*, vol. 195, no. 17, pp. 3834–3844, 2013.

[137] F. S. Berezovskaya, Y. I. Wolf, E. V. Koonin, and G. P. Karev, "Pseudochaotic oscillations in CRISPR-virus coevolution predicted by bifurcation analysis.," *Biology direct*, vol. 9, p. 13, jul 2014.

[138] A. Buckling and P. B. Rainey, "Antagonistic coevolution between a bacterium and a bacteriophage," *Proceedings of the Royal Society B: Biological Sciences*, vol. 269, no. 1494, pp. 931–936, 2002.

[139] K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin, "A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.," *Biology direct*, vol. 1, no. 1, p. 7, 2006.

[140] K. S. Makarova, V. Anantharaman, L. Aravind, and E. V. Koonin, "Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes," *Biology Direct*, vol. 7, no. 1, p. 40, 2012.

[141] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "Comparative genomics of defense systems in archaea and bacteria," *Nucleic Acids Research*, vol. 41, no. 8, pp. 4360–4377, 2013.

[142] Y. Agari, K. Sakamoto, M. Tamakoshi, T. Oshima, S. Kuramitsu, and A. Shinkai, "Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection," *Journal of Molecular Biology*, vol. 395, no. 2, pp. 270–281, 2010.

[143] C. Díez-Villaseñor, C. Almendros, J. García-Martínez, and F. J. M. Mojica, "Diversity of CRISPR loci in Escherichia coli," *Microbiology*, vol. 156, no. 5, pp. 1351–1361, 2010.

[144] I. Grissa, G. Vergnaud, and C. Pourcel, "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats.," *BMC bioinformatics*, vol. 8, p. 172, 2007.

[145] C. Hale, K. Kleppe, R. M. Terns, and M. P. Terns, "Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus.," *RNA (New York, N.Y.)*, vol. 14, no. 12, pp. 2572–9, 2008.

[146] J. Bondy-Denomy and A. R. Davidson, "To acquire or resist: The complex biological effects of CRISPR-Cas systems," *Trends in Microbiology*, vol. 22, no. 4, pp. 218–225, 2014.

[147] S. Bradde, M. Vucelja, T. Tesileanu, and V. Balasubramanian, "Dynamics of adaptive immunity against phage in bacterial populations," *PLOS Computational Biology*, vol. 13, p. e1005486, apr 2017.

[148] C. Díez-Villaseñor, N. M. Guzmán, C. Almendros, J. García-Martínez, and F. J. Mojica, "CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli," *RNA Biology*, vol. 10, no. 5, pp. 792–802, 2013.

[149] S. A. Jackson, R. E. McKenzie, R. D. Fagerlund, S. N. Kieper, P. C. Fineran, and S. J. J. Brouns, "CRISPR-Cas: Adapting to change," *Science*, vol. 356, no. 6333, p. eaal5056, 2017.

[150] J. Zoephel and L. Randau, "RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns," *Biochemical Society Transactions*, vol. 41, pp. 1459–1463, dec 2013.

[151] S. Fischer, L. K. Maier, B. Stoll, J. Brendel, E. Fischer, F. Pfeiffer, M. Dyall-Smith, and A. Marchfelder, "An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA," *Journal of Biological Chemistry*, vol. 287, no. 40, pp. 33351–33365, 2012.

[152] S. Shah, S. Erdmann, F. Mojica, and R. Garrett, "Protospacer recognition motifs," *RNA biology*, vol. 10, no. May, pp. 891–899, 2013.

[153] V. Kunin, R. Sorek, and P. Hugenholtz, "Evolutionary conservation of sequence and secondary structures in CRISPR repeats.," *Genome biology*, vol. 8, no. 4, p. R61, 2007.

[154] K. S. Wilson and P. H. V. Hippel, "Transcription termination at intrinsic terminators: The role of the RNA hairpin (Escherichia coli/RNA polymerase/rho-independent termination)," *Biochemistry*, vol. 92, no. September, pp. 8793–8797, 1995.

[155] P. J. Farnham and T. Platt, "Rho-independent termination: Dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro," *Nucleic Acids Research*, vol. 9, no. 3, pp. 563–577, 1981.

[156] L. Deng, C. S. Kenchappa, X. Peng, Q. She, and R. A. Garrett, "Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus," *Nucleic Acids Research*, vol. 40, no. 6, pp. 2470–2480, 2012.

[157] P. Han and M. W. Deem, "Non-classical phase diagram for virus bacterial coevolution mediated by clustered regularly interspaced short palindromic repeats," *Journal of The Royal Society Interface*, vol. 14, p. 20160905, feb 2017.

[158] S. van Houte, A. Buckling, and E. R. Westra, "Evolutionary Ecology of Prokaryotic Immune Mechanisms," *Microbiology and Molecular Biology Reviews*, vol. 80, pp. 745–763, sep 2016.

[159] P. Payne, L. Geyrhofer, N. H. Barton, and J. P. Bollback, "CRISPR-based herd immunity can limit phage epidemics in bacterial populations," *eLife*, vol. 7, no. 1, p. e32035, 2018.

[160] S. Shmakov, E. Savitskaya, E. Semenova, M. D. Logacheva, K. A. Datsenko, and K. Severinov, "Pervasive generation of oppositely oriented

spacers during CRISPR adaptation," *Nucleic Acids Research*, vol. 42, pp. 5907–5916, may 2014.

[161] E. M. Miller and J. a. Nickoloff, "Escherichia coli Electrotransformation," in *Electroporation Protocols for Microorganisms*, vol. 47, pp. 105–114, New Jersey: Humana Press, 1995.

[162] Anindyajati, A. Anita Artarini, C. Riani, and D. S. Retnoningrum, "Plasmid copy number determination by quantitative polymerase chain reaction," *Scientia Pharmaceutica*, vol. 84, no. 1, pp. 89–101, 2016.

[163] S. Kingsland, "The Refractory Model: The Logistic Curve and the History of Population Ecology," *The Quarterly Review of Biology*, vol. 57, pp. 29–52, mar 1982.

[164] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 1992.

[165] D. Held, K. Yaeger, and R. Novy, "New coexpression vectors for expanded compatibilities in E. coli," *inNovations*, vol. 18, pp. 4–6, 2003.

[166] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003.

[167] K. Hullahalli, M. Rodrigues, U. T. Nguyen, and K. Palmer, "An Attenuated CRISPR-Cas System in Enterococcus faecalis Permits DNA Acquisition," *mBio*, vol. 9, pp. 1–16, may 2018.

[168] C. D. Nadell, K. Drescher, and K. R. Foster, "Spatial structure, cooperation and competition in biofilms," *Nature Reviews Microbiology*, vol. 14, no. 9, pp. 589–600, 2016.

[169] C. D. Nadell, K. R. Foster, and J. B. Xavier, "Emergence of spatial structure in cell groups and the evolution of cooperation," *PLoS Computational Biology*, vol. 6, no. 3, 2010.

[170] I. Frost, W. P. J. Smith, S. Mitri, A. S. Millan, Y. Davit, J. M. Osborne, J. M. Pitt-Francis, R. C. MacLean, and K. R. Foster, "Cooperation, competition and antibiotic resistance in bacterial colonies," *The ISME Journal*, 2018.