The distribution of CRISPR-Cas systems is affected by interactions with DNA repair pathways

Aude Bernheim

UNIVERSITÉ PARIS DESCARTES

École doctorale Frontières du Vivant

Institut Pasteur
Genomique Evolutive des Microbes et Biologie de Synthèse

The distribution of CRISPR-Cas systems is affected by interactions with DNA repair pathways

Par Aude Bernheim

Thèse de doctorat de Microbiologie

Pour obtenir le grade de Docteure de l'Université Paris Descartes

Dirigée par David Bikard, Marie Touchon et Eduardo PC Rocha

Devant un jury composé de :

Olga SOUTOURINA	Professeure - Université Paris-Sud
Olivier TENAILLON	Directeur de Recherches - INSERM
Patrick FORTERRE	Professeur - Institut Pasteur
Bénédicte MICHEL	Directrice de Recherches - CNRS
Edze WESTRA	Associate Professor - Exeter University

La distribution des systèmes CRISPR-Cas est affectée par leurs intéractions avec les systèmes de réparation de l'ADN

Les systèmes CRISPR-Cas confèrent aux bactéries une immunité adaptative contre les éléments génétiques mobiles jouant ainsi un role important dans l'évolution bactérienne. Cependant, moins de la moitié des génomes bactériens encodent des systèmes CRISPR-Cas ; cela, malgré la protection qu'ils confèrent et leur haut taux de transfert horizontal. Des hypothèses telles que le coût des phénomènes d'autoimmunité ou de posséder des défenses adaptatives plutot qu'innées ont été mises en avant pour expliquer ce paradoxe. Je propose une nouvelle hypothèse complémentaire : le contexte génétique jouerait un role important dans la fixation d'un système CRISPR-Cas après son transfert. Plus précisément, j'ai étudié comment les intéractions entre les systèmes de réparation de l'ADN et les CRISPR-Cas influencent la distribution de ces derniers. Pour cela, j'ai d'abord examiné finement la distribution des systèmes CRISPR-Cas dans les génomes bactériens. J'ai ensuite analysé les co-occurences des systèmes de réparation de l'ADN et des CRISPR-Cas et démontré l'existence d'associations positives et négatives entre eux. Enfin, je me suis concentrée sur une des associations négatives découvertes pour valider mes prédictions expérimentalement et comprendre les mécanismes moléculaires sous jacents. Mes travaux permettent de mieux comprendre les intéractions complexes entre systèmes de réparation de l'ADN et CRISPR-Cas et démontrent la necessité d'accomodation des CRISPR-Cas à un contexte génétique pour être selectionnés et maintenus dans les génomes bactériens.

The distribution of CRISPR-Cas systems is affected by interactions with DNA repair pathways

CRISPR-Cas systems confer bacteria and archea an adaptative immunity against phages and other invading genetic elements playing an important role in bacterial evolution. Only 47% of bacterial genomes harbor a CRISPR-Cas system despite their high rate of horizontal transfer. Hypothesis such as the cost of autoimmunity or the trade off between a constitutive or an inducible defense system have been put forward to explain this paradox. I propose that the genetic background plays an important role in the process of maintaining a CRISPR-Cas system after its transfer. More precisely I hypothesized that CRISPR-Cas systems interact with DNA repair pathways. To test this idea, we detected DNA repair pathways and CRISPR-Cas systems in bacterial genomes and studied their co-occurences. We report both positive and negative associations that we interpret as potential antagonistic or synergistic interactions. We then focused on one interaction to validate our result experimentally and explored molecular mechanisms behind those interactions. My findings give insights on the complex interactions between CRISPR-Cas systems and DNA repair mechanisms in bacteria and provide a first example on the necessity of accommodation of CRISPR-Cas systems to a specific genetic context to be selected and maintained in bacterial genomes.

A Alain Bernheim.

Remerciements

La thèse est souvent vue comme une aventure individuelle, mais ce manuscrit est bien pour moi le résultat d'un travail collectif. Tant de personnes à remercier pour m'avoir accompagnée au long de ces trois années alors certes c'est un peu long mais essentiel !

Tout d'abord, mes directeur.rice.s de thèse. J'ai eu la chance de ne pas me limiter à un ni deux mais à trois directeurs de thèse. On m'avait prévenu de possibles complications mais pour moi ce triple encadrement a été une source de richesse importante. Alors dans l'ordre alphabétique des prénoms, merci David pour avoir partagé ton savoir sur les CRISPR, m'avoir initiée à ce que tu as un jour qualifié de "medium throughput biology" ou comment paralléliser un maximum d'expériences en un minimum de temps. Merci d'avoir encouragé ma curiosité en répondant avec patience à toutes mes questions biologiques qu'elles concernent les CRISPR, les phages, les neural networks ou tout autre sujet qui me passait par la tête. Eduardo, j'explique souvent que j'ai pu mesuré mon avancement dans ma thèse au pourcentage de tes remarques que je comprenais du premier coup lors de nos conversations, quoi qu'il arrive ca me faisait souvent réfléchir des jours entiers. J'ai vraiment apprécié toutes les relectures, corrections ... bref le temps que tu as pris pour toujours essayer de me faire progresser. Merci aussi d'avoir cru en moi dès le départ, quand je ne savais ni ce qu'était un plasmide ou ouvrir un terminal en m'acceptant en stage puis en m'expliquant que c'était possible de continuer en thèse en bioinformatique alors que je n'avais aucune formation. Enfin, merci pour ton calme et ta bienveillance qui m'ont bien aidé pendant cette période turbulente qu'a constitué ma thèse. Pour finir, merci Marie de m'avoir à peu près tout appris en bioinformatique: de ma première ligne sous unix aux arbres phylogénétiques en passant biensur par les CRISPR. Merci aussi pour ta patience vis à vis de mes erreurs et errances, j'ai pu apprendre car j'ai pu échouer en sachant que tu serais derrière à m'aider. Enfin, merci pour les multiples conversations qu'on a pu partager aussi bien sur la science que sur tout autre chose.

Alors oui j'ai eu trois directeurs de thèse mais j'ai aussi eu de nombreuses personnes dans mes labos pour me soutenir. Et je ne peux que commencer par Jean. Tu le sais très bien, je n'aurais pas avancé bien loin sans ton aide précieuse et à peu près quotidienne dans tous les domaines. Camille B. et Camille D., mes autres partenaires de bureau, rarement une jungle de bureau n'aura été si joyeuse. Au moins, on savait que dans les mauvais jours, on venait pour voir les membres du 7E. Le nombre de fous rire a du être proportionnel au nombre de problèmes partagés et résolus. Merci Rémi de m'avoir montré tant de choses avec beaucoup de patience. Merci à Olaya, Marc, Pedro, Jorge, Amandine pour les nombreuses discussions et toute l'aide que vous m'avez apportée.

Merci à Clovis et Alicia qui ont travaillé avec moi et ont permis au projet expérimental d'aboutir. Ce fut long, laborieux et plein de rebondissements, mais partager ca avec vous a permis de garder le cap et de toujours se remotiver pour aller plus loin. Merci Florence pour m'avoir montré ce qu'était la rigueur expérimentale. J'ai été ravie de t'avoir comme voisine de paillasse pour te poser toutes mes questions existentielles sur quel erlen utiliser ou qui appeler pour obtenir ce milieu de culture, tout en discutant expos. Merci Antoine et Elise, mes autres co-thésards, Elise pour nos pauses cafés et Antoine pour tes oeuvres d'art contemporaines, nos conversations passionnées sur les oppositions idéalisme pragmatisme et de m'avoir montré pleins de trucs au microscope. Thanks Lun, Gayetri, Belen, Francois for helping me in my work by providing protocols, advice or chinese food!

Pour ce qui est du manuscrit en lui même, merci à mes rapporteurs Olivier Tenaillon et Olga Soutourina pour la relecture de mes travaux. Merci David pour les différentes revisions et Eduardo pour tous ces dimanches à relire ma prose et à la commenter en détails pour me faire aller au bout de mes raisonnements. Vraiment, ce fut un excellent exercice. Merci à Flora pour toutes les questions et correction sur l'introduction. Merci à Jean et Camille D. de m'avoir supportée dans ce bureau pendant ces longues semaines tout en m'aidant avec les choix cornéliens sur les figures, les formulations et les priorités. Et enfin merci à Adrien de m'avoir aidée à essayer de transmettre à un plus large public ce que j'ai pu faire pendant quelques années grâce à la vidéo résumant mes travaux qu'on peut trouver ici => https://vimeo.com/241048776

J'ai aussi la chance d'être entourée d'une communauté de jeunes scientifiques que j'ai pu à multiples reprises consulter pour discuter de mes projets. Merci Antoine D., comme tu le dis si bien, de m'avoir tout appris en biomol et de continuer à debugger mes clonages encore aujourd'hui. Même si je ne le mentionne pas pendant nos Marsouins réguliers, je t'en suis vraiment reconnaissante. Thank you Matt for always always taking the time to chat and share your wonderful ideas about my science. Merci Vincent L. de toutes nos folles conversations, projets communs, critiques constructives et de mettre de temps en temps ton étrange cerveau au service de mes idées. Merci Xavier de m'avoir fait perdre des heures et des heures de ma vie en pauses café! Plus sérieusement, merci pour les nouveaux défis permanents et à travers nos conversations de ne jamais me faire perdre de vue le lien science société.

Il y a des personnes qui rentrent dans plusieurs paragraphes alors je vais en utiliser une pour faire la transition. Car au dela de l'aventure scientifique, j'ai pu dans le cadre de ma thèse explorer de nombreux projets à peu près tous liés au CRI: co-fonder et présider une association sur la promotion des sciences et la mixité des sciences, partir en Chine pour des workshops pour développer des nouveaux masters, co-créer et co-animer un cours à Sciences Po sur les technologies et les politiques publiques ... Ce fut riche. Alors pour commencer, merci Flora d'avoir

partagé à peu près toutes mes aventures. On pouvait switcher de conversations sur les CRISPR aux politiques d'égalité femmes-hommes pendant qu'on se trouvait au château de Chambord pour une intervention avec un verre à la main. Vraiment, pour toute ton aide et notamment la plus récente, la relecture de ma thèse, merci. Merci à Vincent D., Amodsen, Thoma pour des conversations complètement perchées et totalement rafraichissantes. Merci Adrien, Camille C., Lucia, Agathe, Alice, Jerem, Nina et tous les autres pour cette incroyable aventure qu'est WAX Science. Que de moments exceptionnels et vraiment ici il n'y a pas la place. Enfin, merci Francois, Pascal et Ariel de m'avoir si bien fait profiter de cette marque de fabrique du CRI: les mentors. Vous avez toujours pris le temps de discuter, de m'ouvrir l'esprit, de me conseiller.

La thèse c'est aussi trois ans d'une vie. Et dans mon cas, ca a été pour le moins chaotique. Merci donc pour le soutien inestimable que vous m'avez apporté, mes frères et soeurs Adrien et Ségo ; mes frères et soeurs adoptifs du quotidien aka mes collocs Agathe, Camille, Mathieu ; ma fratrie élargie que sont mes cousin.e.s, et mes ami-e-s sans qui je ne m'en sortirais pas l'US, les agros, la team Ogny, les orteaux, les adeptes du Marsouins, les Passy et tous ceux qui ne rentrent dans aucun groupe.

Enfin, je ne peux pas finir ces pages sans mentionner les personnes que je ne peux remercier de vive voix car elles ne sont plus là. Dans ces moments joyeux, il y a toujours une pensée pour elles. En écrivant ces mots je pense notamment à Raphaël, à mes grands-parents et à mes parents. J'ai une pensée toute particulière pour mon père, disparu au début de ma thèse. Homme de sciences, il aurait sans nul doute été heureux de pouvoir partager ce moment avec moi et évidemment de mon côté j'aurais aimé pouvoir lui faire lire mes travaux et en débattre avec lui. C'est pour cela que je lui dédie cette thèse.

Contents

List of Figures

Preamble

For the past ten years, CRISPR-Cas systems have passionated the scientific community both because of their role as an adaptive immune system in bacteria and of their use in many biotechnological applications especially in genome editing. Incredible progress has been made to understand the complex biology and molecular mechanisms of these very diverse systems. However, much remains to be studied on the evolution of CRISPR-Cas systems.

One intriguing observation is that only 50% of bacterial genomes harbor a CRISPR-Cas system [161] despite their apparent fitness advantage and their high rate of horizontal transfer. The focus of my thesis was to tackle this question by understanding what could be the evolutionary downsides of CRISPR-Cas systems. Hypotheses such as the cost of autoimmunity or the trade off between a constitutive versus a specialized defense system have been put forward to explain this paradox. I propose a new and complementary hypothesis: that the genetic background plays an important role in the process of fixation of a CRISPR-Cas system acquired by horizontal transfer.

To test this hypothesis, we decided to focus on one essential bacterial function: DNA repair. Our choice was motivated by several reasons. First, CRISPR-Cas systems and DNA repair pathways share the same substrate, DNA. Second, by potentially being able to repair breaks generated by CRISPR-Cas systems, DNA repair pathways could limit CRISPR-Cas efficiency. Therefore, the goals of my thesis were to 1) characterize precisely CRISPR-Cas distribution 2) study if DNA repair pathways interactions with CRISPR-Cas systems impacted the distribution of CRISPR-Cas systems in bacterial genomes.

Chapter 1 introduces CRISPR-Cas systems, presents briefly DNA repair pathways and examines the importance of known interactions between CRISPR-Cas systems and DNA repair pathways. To answer my PhD questions, I used both bioinformatics and molecular biology. Therefore, in Chapter 2, I introduce methods to study CRISPR-Cas systems from a bioinformatics and experimental point of view. In Chapter 3, I present a description of CRISPR-Cas systems distribution, organization, interactions and transfers that represent the first integrated analysis of Cas operons and CRISPR arrays. In Chapter 4, I introduced evidence that DNA

repair pathways interactions with CRISPR-Cas systems shape CRISPR-Cas distribution in bacterial genomes by studying co-occurrence patterns of DNA repair pathways and CRISPR-Cas systems. In Chapter 5, I confirm experimentally the importance of those interactions by exploring the molecular mechanisms behind one of the proposed interactions: an antagonism between the type II-A CRISPR-Cas system and the Non Homologous End Joining Pathway (NHEJ). Finally, in Chapter 6, I discuss the relevance of these findings and open questions about CRISPR-Cas systems evolution and biology.

Chapter 3, 4 and 5 correspond to three manuscripts for which I was the main contributor. Two of them are in preparation (3 and 4) and one has just been accepted for publication in *Nature Communications* (5). Chapter 2 is partly based on a book chapter that I co-authored. Annexes 2 and 3 present two articles to which I contributed during my PhD but that are not related directly to my PhD subject. The first one (Annexe 2) studies the determinants of the distribution of prophages in bacterial genomes and was published in ISME journal. The second one (Annexe 3), reports the discovery of a new defense pathway in *Staphylococci* and was published in Cell Host and Microbe.

Chapter 1

Introduction

1.1 CRISPR-Cas systems fundamentals

1.1.1 A (very) small history of CRISPR-Cas systems discovery

Over the past years, several articles and books told the story of CRISPR-Cas systems discovery [146, 20, 66]. They can be read as short novels of how does a scientific breakthrough take place in the 21st century. Without paraphrasing such passionating reads, I will give the main milestones of the CRISPR field which are summarized in Figure 1.1.

CRISPR arrays were first described in 1987 in a bioinformatic study of *Escherichia coli* [121]. In 2002, the name CRISPR was coined and *cas* genes were described [125, 126]. In 2005, several teams reported that spacers derive from foreign genetic elements [182, 217, 34], and a year later the function of an adaptive immune system was hypothesized [160]. In 2007, a now seminal paper provided experimental evidence that CRISPR-Cas systems constitute an adaptive immune system against phages [19]. From then on, many more teams contributed to the understanding of the diverse molecular mechanisms at play behind this adaptive immunity. In 2012, the precise mechanism by which type II CRISPR-Cas systems achieve immunity was understood and Cas9 proposed as a reprogrammable RNA guided nuclease suitable for genome editing [88, 130]. In 2013, the first Cas9-based genome editing of human cells [53, 165] and bacteria [127] are reported, paving the way for genome editing of many more organisms in the following years.

Since then, the pace of new discoveries in the CRISPR field has neither slowed down on the fundamental understanding of how CRISPR-Cas systems work nor on the engineering of diverse CRISPR-based technologies. Since, I started my PhD in 2014, the number of scientific publications mentioning the term CRISPR as counted in Pubmed has risen from 610 in 2014 to 2077 in 2016 and is already at

2313 for 2017. In the course of writing the introduction of this thesis, I updated the molecular mechanism section twice on two major points reflecting the speed at which knowledge is produced. I will therefore try to give an up-to-date view of the biology of CRISPR-Cas systems.

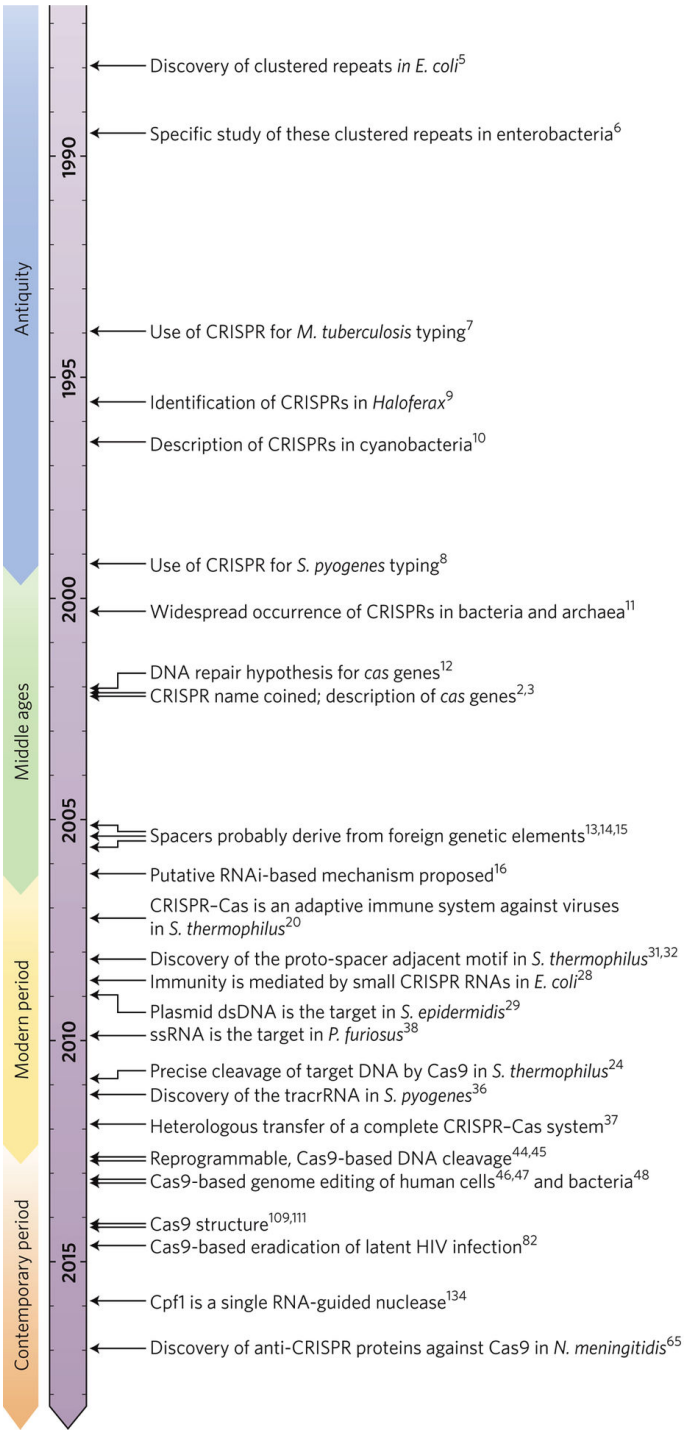


Figure 1.1: CRISPR milestones and seminal discoveries. From [20]

1.1.2 CRISPR-Cas systems : a general description

CRISPR-Cas systems are an adaptive immune system of bacteria and archaea targeting mobile genetics elements (MGE) such as phages or plasmids [19, 40, 169, 87]. They are composed of a CRISPR array (Clustered Regularly Interspaced Spacer withPalindromic Repeats) and an cluster of *cas* genes (CRISPR associated genes) (Figure 1.2). CRISPR arrays comprise two types of sequences: repeats and spacers. Repeats are short sequences (typically 20-40 bp) identical in a given CRISPR array. They are interspaced by short and diverse spacer sequences (typically 20-40 bp), which often match sequences from mobile genetic elements. The leader region of the array contains the start site of transcription of the CRISPR array. The second element of CRISPR-Cas systems consists of a set of *cas* genes necessary to achieve immunity [182, 34, 217, 160].

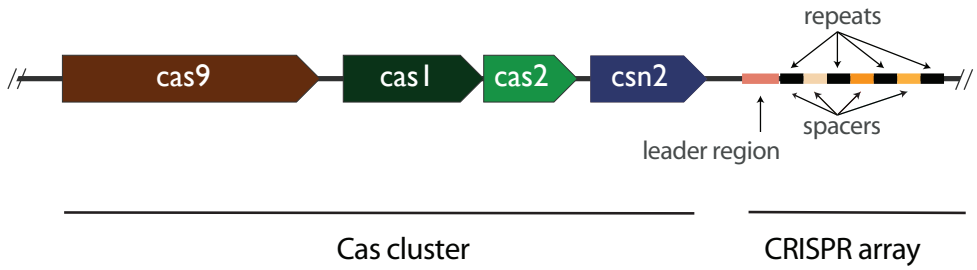


Figure 1.2: Example of a CRISPR-Cas locus.
A CRISPR-Cas locus is organized around two main elements: a cluster of *cas* genes and a CRISPR array composed of spacers and repeats.

CRISPR-Cas immunity works in two stages: immunity and immunization also called adaptation (Figure 1.3) [168]. Immunity can be split in two phases: expression and maturation of crRNA and interference. The CRISPR array is transcribed and then processed into smaller RNAs, each composed of a repeat and a single spacer called crRNA (CRISPR RNA). Each of these crRNA then serves as a guide for a complex of Cas proteins. If the sequence of a guide is identical to another sequence of DNA in the cell like for example DNA from a phage, the complex will activate an immune response most of the time by cutting the invading DNA which will then be processed and degraded. During adaptation, a complex of Cas proteins generates and then incorporates a new spacer in the CRISPR array [168]. An important aspect of the system is the ability to distinguish its own DNA from exogenous one. In many cases, a short sequence named the PAM (Protospacer Adjacent Motif) is essential for this process. Present near a protosopacer (sequence of a spacer on a mobile genetic element) but absent in the CRISPR array, the PAM is necessary for the interference to take place [39].

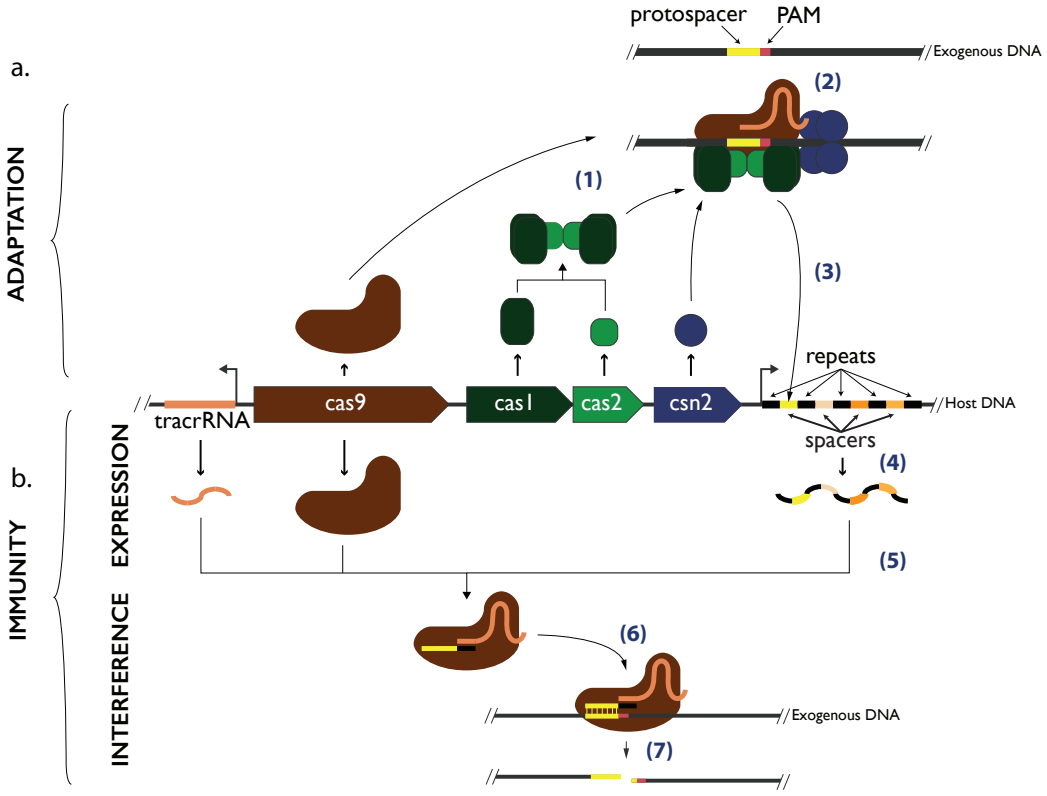


Figure 1.3: Overview of CRISPR-Cas mechanism.
CRISPR-Cas immunity works in two main phases: adaptation where the complex of Cas proteins allows the acquisition of new spacers and immunity where the system provides targeted immunity . **a.** During the adaptation phase, Cas proteins form a complex (1) that is able to generate a short DNA fragment from foreign DNA present in the cell (yellow) (2) that is then integrated in the CRISPR array as a new spacer (3). **b.** During the phase of immunity, the CRISPR array is transcribed (4) and processed into small RNAs (5) which serve as guide for target recognition for Cas proteins (6), which upon match degrade DNA (7).

1.1.3 The diversity of CRISPR-Cas systems

CRISPR-Cas systems are extremely diverse. There are classified in two classes, six types (I to VI) and 21 subtypes [161, 181, 140, 163]. New subtypes are discovered every year leading to an evolving classification [244, 43] (Figure 1.4.a). Classification of CRISPR-Cas systems is based on the architecture of the loci and the content of the Cas cluster, more specifically signature proteins [161] : Cas3 for type I, Cas9 for type II, Cas10 for type III, Csf4 for type IV, Cas12 for type V and Cas13 for type VI. One key aspect of CRISPR-Cas systems organization is their modularity [181, 161] (Figure 1.4.b). The diverse CRISPR-Cas systems are organized around four modules: crRNA processing, target recognition, target cleavage and adaptation. While the adaptation module is largely similar in all CRISPR-

Cas systems and involves Cas1 and Cas2, the other modules can encompass a very small or large number of proteins [161, 181].

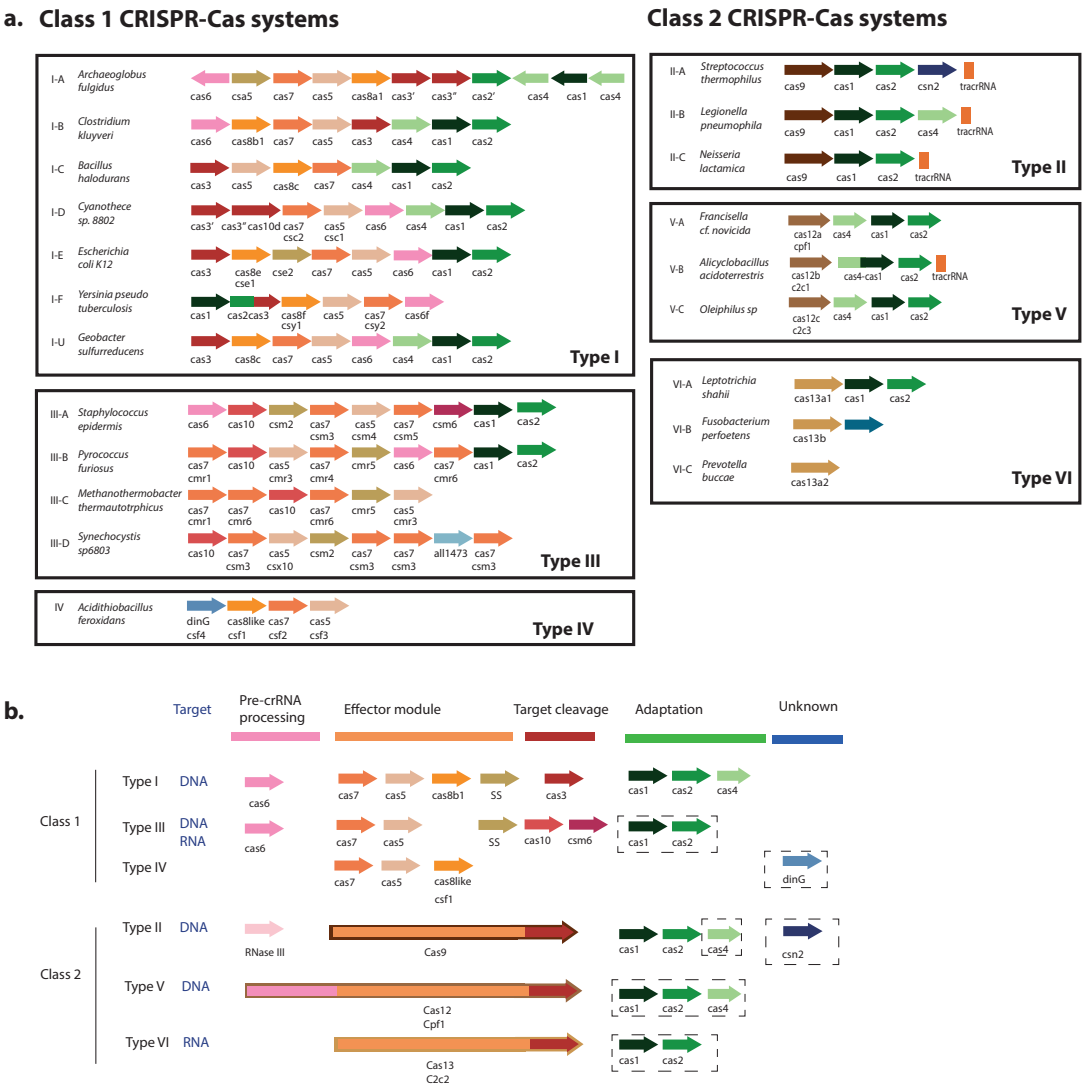


Figure 1.4: The diversity of CRISPR-Cas systems
a. Classification of CRISPR-Cas systems. CRISPR-Cas systems are classified in two classes, six types (I to VI) and 21 subtypes. Classification is based on signature proteins and architecture of the loci. Organisms harbouring a subtype are presented next to the subtype name. **b.** The modular organization of CRISPR-Cas systems. CRISPR-Cas systems are organized around four modules: crRNA processing, target recognition, target cleavage and adaptation. Proteins that are absent in specific subtypes but present in others are surrounded by dashed lines. The colours of the arrows representing proteins correspond to the module to which they belong. Adapted from [161, 181].

Class 1 CRISPR-Cas systems are classified into three types and 12 subtypes [161] (Figure 4.a). This class of systems is defined by its use of multi-protein effector complexes, that is to say that the interference phase is carried out by a complex of multiple proteins. They are by far the most abundant as they represent 90% of the CRISPR-Cas systems [161]. On the other hand, Class 2 CRISPR-Cas systems only represent 10% of CRISPR-Cas systems and are almost completely absent from archaea [161, 42]. Their effector complex is represented by a single multidomains protein. CRISPR-Cas based technologies are derived from subtypes belonging to this Class.

CRISPR-Cas systems are also diverse with respect to their architectures. While most of the time *cas* genes are found near CRISPR arrays, orphan arrays (ie CRISPR array without neighboring *cas* genes) have been frequently identified using bioinformatics [161, 244]. They can be processed by *cas* genes present *in trans* or inactive [19].

The diversity of Cas proteins and architecture loci reflect a diversity of molecular mechanisms by which CRISPR-Cas systems achieve adaptive immunity.

1.1.4 Molecular mechanisms of immunity

Class 1 CRISPR-Cas systems

Class 1 CRISPR-Cas systems require multi-proteins effector complexes to cleave a target nucleic acid sequence. While type I systems exclusively target and cleave DNA, type III systems can recognize and cleave DNA and RNA [168]. The following description of molecular mechanisms of the immunity phase details 1) the expression and processing of the long precursor RNA transcribed from the CRISPR array into mature crRNA 2) the loading of the crRNA on the effector complex which then recognizes the specific target nucleic sequence 3) the subsequent cleavage of the target. While in most cases the target sequence comes from a foreign piece of DNA such as a phage or a plasmid, autoimmunity ie targeting of bacteria's own chromosome has been reported for all types of CRISPR-Cas systems [111]. Molecular details are summarized in Figure 1.5.

Type I and III CRISPR-Cas systems immune response starts with the transcription of the CRISPR array into a long precursor RNA (Figure 1.5.a). This precursor RNA is then processed by Cas6, an endoribonuclease, to produce mature crRNA [45, 190] except in type I-C where this function is performed by Cas5 [187]. A mature crRNA consists of a single spacer flanked by a repeat [40]. In type I systems, Cas6 is part of the effector complex while it acts without the complex in type III [298, 158, 257]. Once processed, the crRNA is loaded on the effector complex, refered to as Cascade for type I, Csm/Cmr complex for type III (Figure 1.5.b). The effector complex is organized in a similar manner for type I and type

III systems [255, 229, 314]. It is composed mainly of proteins from the RAMP (repeat associated mysterious proteins) family like Cas5 and Cas7. Typically, several Cas7 subunits interact with the small subunit (Csa5 or Cse2 for type I and Csm2 or Cmr5 for type III) and bind the crRNA backbone. One Cas5 subunit binds to the 5 end of the crRNA and interacts with the large subunit (Cas8 for type I and Cas10 for type III) [132, 224, 230, 255, 229, 314] (Figure 1.5).

In type I systems, upon recognition of a PAM, an R-loop (three-stranded nucleic acid structure, composed of a DNA:RNA hybrid and the non-template single-stranded DNA) is formed between the crRNA and the target dsDNA (Figure 1.5.c). The 8 base pairs at the 5' end of the crRNA, called the seed, are particularly important as mutations in this region lead to loss of immunity [238]. Once the R-loop is formed, the Cas3 helicase-nuclease is recruited by the effector complex and introduces single strand DNA breaks (Figure 1.5.d) [40, 33, 186, 123] .

In type III systems, once loaded on the effector complex, the crRNA undergoes further maturation to trim its 3' end. Cas10 only cleaves target DNA when it undergoes transcription (Figure 1.5.e) [92, 232, 267]. Robust interference involves cleavage of both DNA and the nascent mRNA [92, 232, 78, 129].The mechanism of self/non-self discrimination does not involve a PAM. It is based on the complementarity between a part of the crRNA known as the crRNA tag and the sequence flanking the protospacer. DNA targeting will only happen when the pairing is imperfect, as perfect pairing between the crRNA tag and the repeat prevents autoimmunity [170]. For some type III systems, in presence of a target DNA, Cas10 will not only cleave DNA but synthesizes a small molecule called cyclic oligoadenylates cOA, which will activate Csm6, a RNase that will then cleave RNA in a non specific manner (Figure 1.5.e) [136, 189]. The accumulation of cOA constitutes an intracellular signal that infection has not been prevented, as the CRISPR-Cas systems is still active and leads to cell death or dormancy to prevent further propagation of the phage [136, 189, 5]. Some type III systems (III-A and III-B) can also target foreign RNA. RNA is then cleaved by specific proteins of the Cas7 family (Csm3 and Cmr4) (Figure 1.5.f) [256, 24, 220].

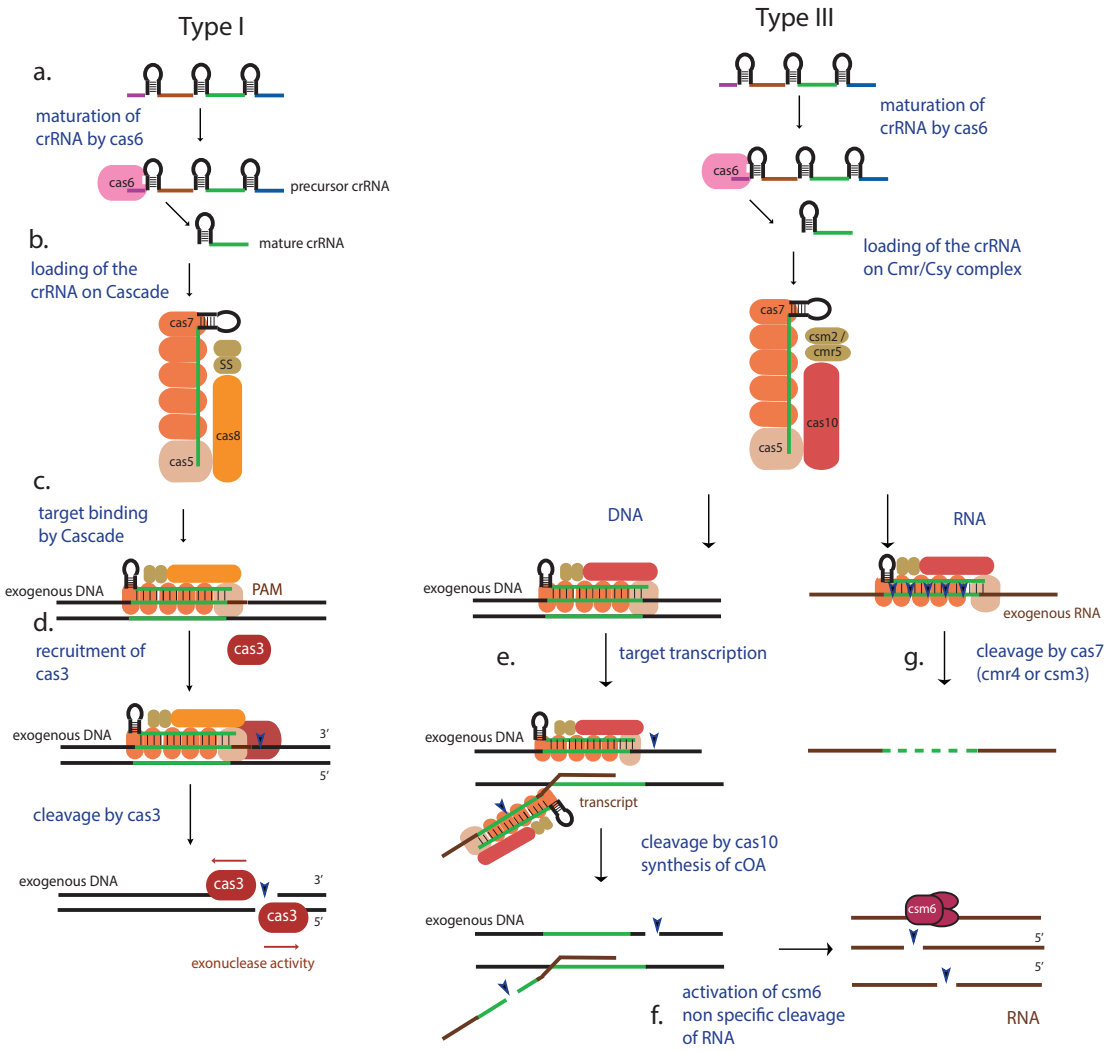


Figure 1.5: Class 1 CRISPR-Cas systems molecular mechanisms of immunity **a.** CRISPR array is transcribed into a long precursor RNA and processed by Cas6 (in pink) to produce mature crRNA ie single spacers flanked by a repeat. **b.** crRNA is loaded on the effector complex refered to as Cascade for type I, Csm/Cmr complex for type III. Typically, several Cas7 subunits (in orange) interact with the small subunit (in beige) and bind the crRNA. **c.** In type I systems, an R-loop (dsDNA:RNA hybrid) is formed upon PAM (in brown) recognition by Cascade. **d.** This leads to the recruitment of the nuclease Cas3 (in red) which introduces single strand DNA breaks. **e.** In type III systems, Cas10 cleaves DNA in presence of a transcript. **f.** Cas10 also synthetizes a molecule that activates a non-specific RNase Csm6. **g.** Subtypes III-A and III-B systems can also cleave RNA.
Adapted from [181, 164].

Class 2 CRISPR-Cas systems

Class 2 CRISPR-Cas systems require a single effector protein to cleave a target nucleic acid sequence. Type II and type V act on dsDNA while type VI cleaves

ssRNA [130, 61, 312, 82].

Type II systems depend on two small RNAs for immunity: the tracrRNA (trans-activating crRNA) and the crRNA. The tracrRNA possess a region complementary to the repeat of the CRISPR array. When in association with the transcribed CRISPR array, the CRISPR primary transcript and the tracrRNA are processed by the RNAse III (a host protein) at each repeat generating single crRNA guides (Figure 1.6.a) [130, 61]. In the type II-C system from *Neisseria meningitidis*, intermediate crRNA guides are transcribed from multiple promoters embedded within the repeats of the CRISPR array [316]. Type V-B CRISPR-Cas systems also rely on a tracrRNA while type V-A and type VI do not. Cas12a from the Type V-A systems has an intrinsic RNAse activity enabling it to process the pre-crRNA without requirements for host factors (Figure 1.5) [312, 82, 244, 2, 69].

In type II systems, duplexes of crRNA and tracrRNA are then loaded on Cas9 to form the interference complex (Figure 1.6.b). Type II immunity requires a PAM located downstream of the target sequence. Cas9 binds transiently to PAM sequences and probes the complementarity of the first 6-8bp of the crRNA with the target sequence. A good complementarity will trigger the formation of an R-loop which will set off a conformational change of the nuclease domains of Cas9 (RuvC and HNH) leading to the cleavage of each strand of the target generating a blunt end double strand break (Figure 1.6.c) [130, 88, 262, 131, 191, 7, 261].

Like type II systems, type V effectors encompass a RuvC like domain. They require a T-rich PAM at the 5 of the protospacer. While type V-B effector functions in a similar manner as Cas9 with a tracrRNA, type V-A cleaves in a stepwise manner. A RuvC-domain-dependent cleavage is followed by an allosterical change leading to a second cleavage by another nuclease domain, generating a staggered double strand break with 4-5 bp overhang (Figure 1.6.c) [242, 312, 82, 305].

In contrast to other effector complexs of Class 2, type VI systems do not possess a RuvC domain but two HEPN domains. The signature protein of these systems, Cas13a (C2c2), is a RNA-guided RNAse. Upon recognition of the target RNA, Cas13a changes into a non specific RNase leading to cell toxicity and/or death (Figure. 1.6.c) [242, 36, 2, 69].

The molecular mechanisms of CRISPR-Cas immunity are extremely diverse. While most of the details for certain subtypes have been completely elucidated, some remain to be discovered such as the mechanisms of type IV CRISPR-Cas immunity.

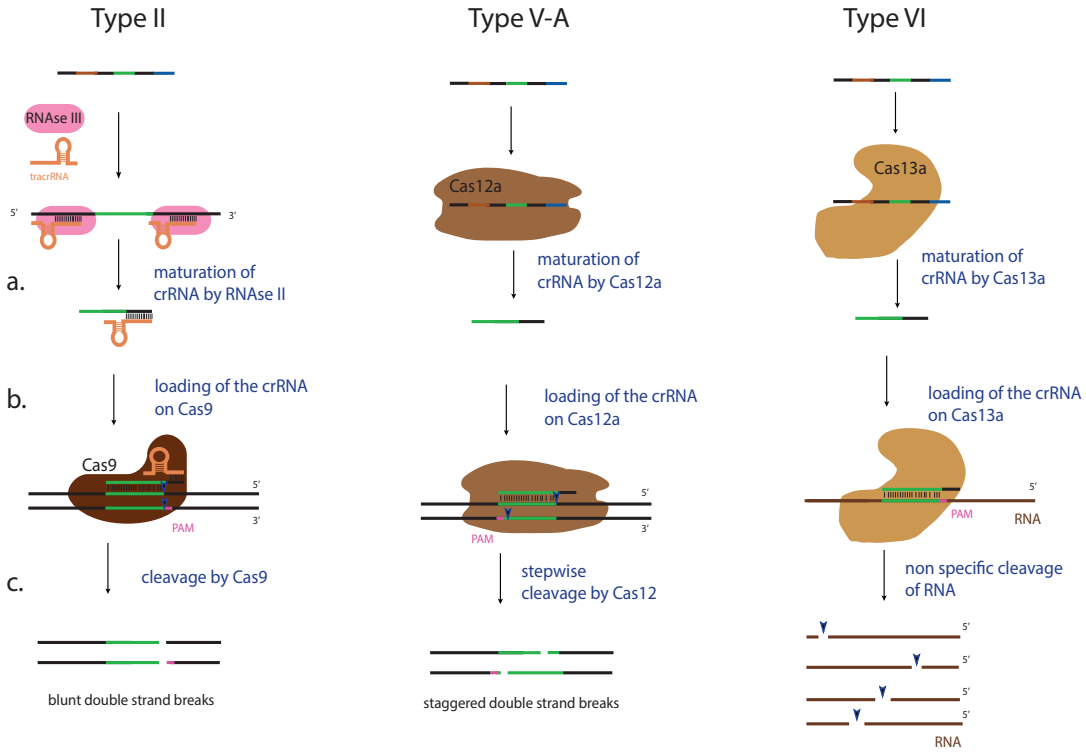


Figure 1.6: Class 2 CRISPR-Cas systems mechanisms of immunity.
a. Type II systems immunity relies on the tracrRNA (orange) and the crRNA (black =repeat , green =spacer). The crRNA in association with the tracrRNA is processed by RNase III (pink). In type V and VI, crRNA is processed by Cas12a, Cas13a. **b.** crRNA is then loaded on Cas9, Cas12a or Cas13a to form the interference complex. Upon PAM (pink) recognition and sequence complementarity, an R-loop (dsDNA:RNA hybrid) is formed for Cas9 and Cas12a while Cas13a recognizes and targetd RNA. **c.** Conformational change activates nuclease domains. In type II systems, Cas9 generates double strand breaks. In type V-A systems, Cas12a cleaves in a stepwise manner generating staggered break. In type VI systems, non specific RNA cleavage is activated upon RNA recognition. Adapted from [181, 164].

1.1.5 Molecular mechanisms of adaptation

In contrast with the diversity of proteins involved in immunity, the adaptation phase is mainly carried out by two proteins: Cas1 and Cas2. Cas1 and Cas2 are present in most CRISPR-Cas systems. It is believed that subtypes which lack Cas1 and Cas2 might acquire new spacers through the use of Cas1 and Cas2 present *in trans* in the bacterial genome. The process of adaptation requires several steps: 1) production of prespacers *i.e.* small pieces of DNA which will be processed and integrated in the CRISPR array 2) recognition of the CRISPR array by the adaptation machinery to integrate the new spacer at the right position 3) integration of the new spacer in the CRISPR array. Even if Cas1 and Cas2 are central to all CRISPR-Cas systems, some aspects of the spacer acquisition mechanism remain

specific to types or subtypes like prespacers production or primed adaptation [124].

Production of prespacers from naïve adaptation

The first step of adaptation is the production of prespacers which corresponds to the generation and the selection of pieces of DNA that will then be loaded on the adaptation machinery (the Cas1-Cas2 complex) and further integrated in the CRISPR array. There are two main pathways to generate new spacers: naïve or primed adaptation. Naïve adaptation can be defined as the acquisition of a spacer from a mobile genetic element (MGE) against which no other spacer in the host genome exists. On the other hand, primed adaptation is the acquisition of a spacer from an MGE already targeted by a spacer present in the bacterial genome.

For naïve adaptation, today, only one main route of prespacers production has been identified for type I-E systems in *E. coli* [149]. Cas1-Cas2 would take the degradation products of RecBCD activity (a DNA repair complex) as DNA substrate. As an important number of dsDNA breaks in the cell occurs during DNA replication, plasmids present in many copies in the cell would be more prone to spacer acquisition. The high density presence of Chi sites on the bacterial chromosome would further protect it from spacer acquisition (Figure 1.7.a) [149]. However, naïve adaptation still takes place in *E. coli* without RecBCD proving the existence of other routes (Figure 1.7) [149]. Other potential prespacers could derive from byproducts of other cellular machineries. For example, fragments generated by restriction-modification systems could serve as prespacers for CRISPR-Cas systems linking innate and adaptive immunity (Figure 1.7.a) [68].

Another important aspect of spacer selection, is the presence of a correct PAM on the targeted MGE. Part of this PAM selection has been elucidated for type II systems (Figure 1.7.b). It was first shown that all the proteins of type II-A systems are necessary to achieve adaptation [287]. Cas9 plays a specific role in PAM selection as changing the Cas9 PAM specificity for example from NGG to NGGNG changes the selection of PAMs in prespacer substrates [108]. However, the role of Cas9 in adaptation does not seem limited to PAM selection as specific mutations in the protein Cas9 can lead to variants that increase the rate of spacer acquisition up to a 100 fold [109].

The timeline of events is important to achieve immunity to a novel element as CRISPR-Cas systems need to acquire a new spacer and perform interference before the phage completes its replication cycle. Two studies have brought elements to clarify this issue. First, CRISPR-Cas systems acquire spacers from defective phages (Figure 1.7.b) [119]. More recently, a study established that spacers are acquired at the beginning of the infection during DNA injection [179]. Spacers acquired during that phase are more effective at degrading invading DNA during interference than spacers that would be acquired during other phases of the phage

life cycle. This contributes to explain how CRISPR-Cas systems generate spacers that will be successful in achieving efficient immunity (Figure 1.7.b) [179].

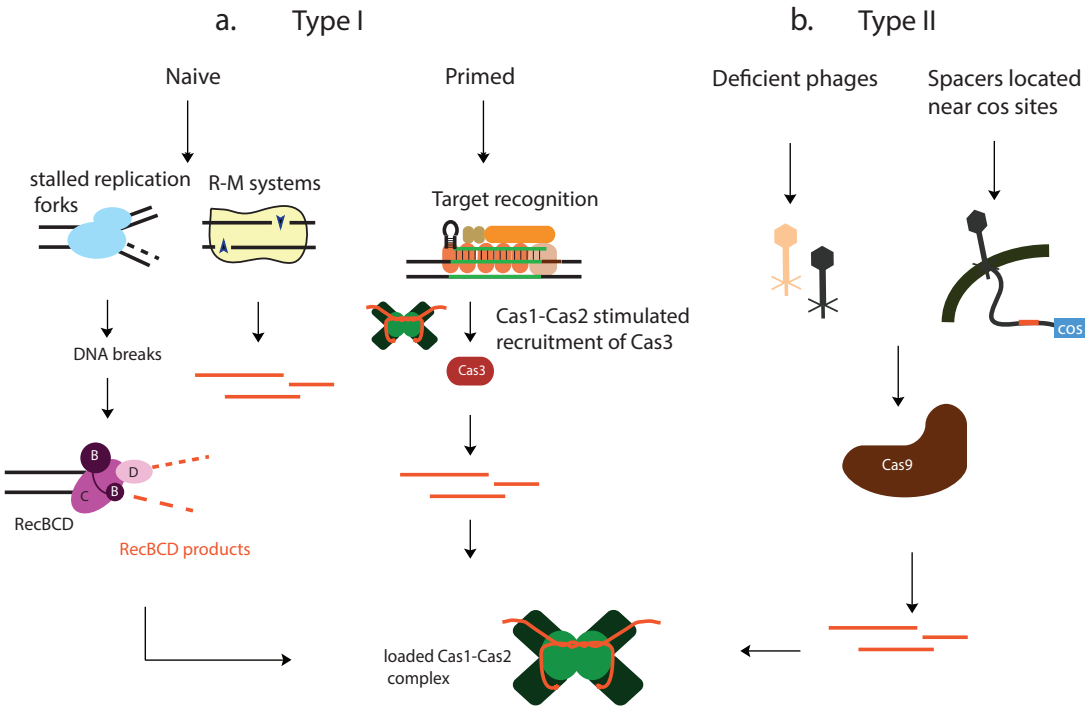


Figure 1.7: Prespacers production pathways.
a. Prespacers (red) can either be produced by naïve or primed adaptation by type I systems. RecBCD is involved in prespacers production in naïve adaptation of type I-E systems. R-M systems could also generate prespacers in naïve adaptation. Primed adaptation occurs through a Cas1-Cas2 stimulated recruitment of Cas3 which generates small DNA fragments which can serve as prespacers. **b.** In type II systems, prespacers can be acquired from defective phages or near cos sites *i.e.* DNA that is first injected into the bacteria. Cas9 plays an important role in PAM selection. Adapted from [124].

Finally, another original way of acquiring new spacers was uncovered in some type III systems where Cas1 is fused to a reverse transcriptase (RT-Cas1) [160, 268]. RT-Cas1 allows the direct incorporation of RNA spacers. RT-Cas1 associated with Cas2 catalyzes the ligation of RNA segments into the CRISPR array which is then followed by reverse transcription and replacement of the RNA strand by DNA [249]. This process also ensures that novel spacers come from transcribed regions and are integrated in the right orientation, a prerequisite for type III immunity.

Production of pre-spacers from primed adaptation

MGE can easily escape CRISPR-Cas immunity by mutating one position. In order to keep up with escape mutants, type I systems can acquire spacers through primed adaptation [265, 150, 58, 226]. Primed adaptation allows spacers acquisition at a higher rate from an MGE already targeted by a spacer than by naïve adaptation (Figure 1.7.a)[81].

Primed adaptation is based on an imperfect target recognition and therefore requires both the machinery of adaptation and interference [281]. Molecular mechanism requires the effector complex to choose between interference or primed adaptation. The first step is target recognition. The detection of a PAM or a specific crRNA sequence can induce conformational rearrangements in the bound crRNA effector complex that will favor either priming or interference [224, 303, 104, 304, 302, 275]. In type I-E systems, Cas8, the large subunit of the effector complex can adopt two conformational modes which will either lead to a direct recruitment of the Cas3 nuclease and therefore interference [104, 304] or a Cas1-Cas2 stimulated recruitment of Cas3 [224, 302]. In the case of a Cas1-Cas2 recruitment, Cas3 of type I-E produces in vitro ssDNA fragments enriched for PAMs [144]. They could provide prespacers for Cas1-Cas2 which could be localized near Cas3 *i.e.* the prespacers production site. In type I-F, *cas3* is fused to *cas2* [161, 226, 258] leading to a colocalization of the prespacer production site and the adaptation machinery.

As primed adaption can still function with guides with several mismatches, it constitutes a positive feedback loop against escape mutants [124]. This strategy can be essential to achieve complete immunity against specific MGE such as highly mutating phages [150].

Integration of a new spacer

Once a prespacer is loaded on the Cas1-Cas2 integration complex, integration of a new spacer will unfold in three steps: 1) ensuring the right orientation and positioning of the spacer, 2) finding the right position to integrate the spacer, 3) integrating the spacer (Figure 1.8).

First, to determine the right size, type I-E Cas1-Cas2 complex uses tyrosine wedges present on the Cas1 dimer to act as a molecular ruler. The length of the new spacer is therefore determined by the fixed distance between those wedges (Figure 1.8.a) [282, 194, 95]. Given that spacers in one array have usually the exact same size, similar molecular rulers could exist in other systems. Correct PAM selection is still not fully understood but in type I-E systems, the presence of a canonical PAM within the prespacer increases the affinity for Cas1-Cas2 binding even if is not a requisite [282].

The second step is the recognition of the CRISPR array (Figure 1.8.b). The Cas1-Cas2 prespacer complex binds to the leader region of the CRISPR array and first repeat. To do so, it is either directed by sequences upstream of the leader region and/or assisted by host proteins. Cas1-Cas2 complex shows specific affinity in vitro for sequences upstream of the leader region [193, 301, 173, 284]. In type I-F systems, the leader repeat recognition is assisted by the integration host factor (IHF) heterodimer [192, 301]. IHF binds the CRISPR leader to induce DNA bending helping Cas1-Cas2 localize the leader region [308]. However, IHF is not present in all bacteria, pointing out the existence of other mechanisms to specify the integration site. In type II systems, short leader-anchoring sites adjacent to the first repeat are essential for adaptation. Therefore type II-A relies only on sequence specificity for the leader repeat recognition [287, 193, 173, 302].

The third step is the integration into the CRISPR array (Figure 1.8.c). For type I-E systems, the four Cas1 monomers contain a PAM-sensing domain, only one PAM sensing domain is sufficient to appropriately place the substrate in the right orientation for integration [282, 194, 265, 243]. Orientation is critical; if the PAM ends up on the wrong side of the protospacer on the targeted MGE, interference will not take place. Processing of the prespacer by Cas2 creates two 3'OH ends required for nucleophilic attack on each strand of the leader-proximal repeat [13, 193]. Two consecutive nucleophilic attacks generate the full site integration product. The first nucleophilic attack most likely occurs at the leader-repeat junction and creates a half-site intermediate. The second nucleophilic attack occurs at the junction between the repeat and the spacer. Once the full site integration product is created, host DNA repair enzymes fill the gaps generating new repeats (Figure 1.8.d) [193, 301, 124].

Spacer integration mechanisms of type I-E and type II-A are now relatively well understood. Several differences can be noted such as the mechanism of identification of the leader end of the CRISPR array through specific sequence for type II-A [302] and with the help of IHF for type I-E [192, 301]. Other mechanisms could exist for other subtypes and remain to be discovered.

To conclude, molecular mechanisms by which CRISPR-Cas systems acquire new spacers and target specific nucleic acid sequences are extremely diverse. Their study has led to the emergence of a multitude of CRISPR based technologies and the understanding of the rest of these mechanisms will likely enrich the CRISPR-Cas toolbok. The study of specific molecular details for certain subtypes was hampered by the initial observation that CRISPR-Cas systems were not always expressed. The following section therefore details CRISPR-Cas systems regulation.

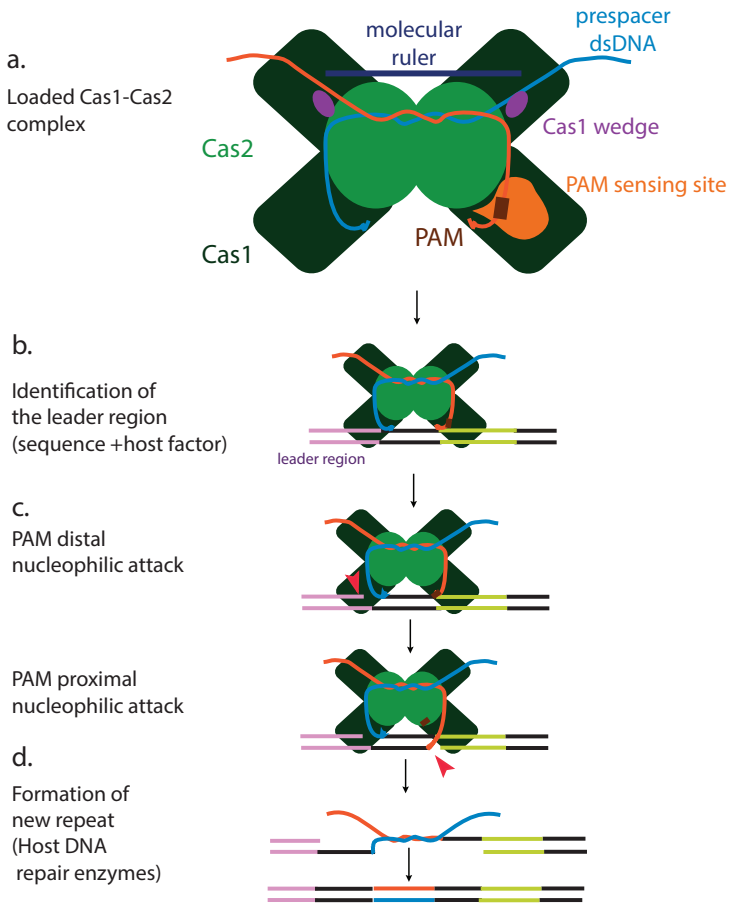


Figure 1.8: Spacer integration.
a. Loaded Cas-Cas2 complex (green). Correct size of the spacer is guaranteed by the distance between tyrosine wedges on Cas1 (purple) that act as molecular ruler (dark blue). Cas1 also possess a PAM sensing site which ensure correct orientation (orange). **b.** Leader end recognition is achieved by sequence specificity or helped by host factors like the IHF (not represented) that bends the DNA. **c.** A new spacer is integrated through two consecutive nucleophilic attacks (red arrows) of bound DNA to the CRISPR array **d.** New repeat is formed by gap filling by host DNA repair proteins. Red and blue lines correspond to two strands of the loaded prespacer. Adapted from [124].

1.1.6 CRISPR-Cas systems regulation

CRISPR-Cas systems are regulated by diverse mechanisms [206]. The most famous regulation of CRISPR-Cas systems comes from *E. coli*, as its endogenous type I-E CRISPR-Cas system is not expressed under laboratory conditions [295]. The type I-E CRISPR-Cas system is repressed by the histone-like nucleoid-structuring (H-NS) which can be relieved by the transcription activator LeuO (Figure 1.9.a)[295]. This type of repression is actually not specific to *E. coli* but shared by *Salmonella enterica* serovar Typhi and *Klebsiella pneumoniae* [174, 151]. In *K.*

pneumoniae, imipenem antibiotic increases H-NS expression, which then decreases *cas3* transcription, and leads to reduced CRISPR-Cas activity [151]. In *Salmonella enterica* serovar Typhi, the leucine-responsive regulatory protein (LRP) is involved in CRISPR-Cas repression (Figure 1.9.a) [174]. This endogenous repression could limit the cost of expressing constitutively a CRISPR-Cas system or limit autoimmunity phenomenon. One major aspect of CRISPR-Cas systems regulation is therefore to understand the external factors that will lead to the upregulation of CRISPR-Cas systems compared to a repressed state.

First, some CRISPR-Cas systems are regulated by quorum sensing (Figure 1.9.b) [205, 116]. At low density, the quorum sensing machinery of *Serratia sp.* ATCC39006 represses three different CRISPR-Cas systems (I-E, I-F and III-A). The accumulation of quorum sensing signal in high density population leads to an increased expression of the CRISPR-Cas systems [205]. In *Pseudomonas aeruginosa*, type I-F CRISPR-Cas system is controlled by two quorum sensing activators [116]. In both cases, the argument for the existence of this regulation, is the ability of the cell to limit the cost of expressing its defense systems at low density when the risk of infection is lower compared to high density (Figure 1.9) [205, 116].

Second, several arguments point to the ability of CRISPR-Cas systems to respond to the disruption of the bacterial envelope (Figure 1.9.c) [222]. First, it was observed in different organisms that CRISPR-Cas are upregulated during phage infection [310, 219, 84, 3]. Second, in *E. coli*, a link was demonstrated between the type I-E CRISPR-Cas system and the regulator system BaeSR known to respond to extracytoplasmic stress [211]. Moreover, type I-E CRISPR-Cas system is upregulated through the stabilization of Cas3 by the heat shock protein G (HtpG), known to be induced during phage infection [309]. Finally, a third potential link between CRISPR-Cas systems and the bacterial envelop resides in the VicRK system of *Streptococcus mutans*, which is implicated in response to oxidative stress, competence and biofilm formation. Mutants deleted for VicRK resulted in increased type II-A expression [239].

Third, CRISPR-Cas systems are regulated by metabolic changes which could signal an infection (Figure 1.9.d). The argument relies on the observation that bacterial metabolism changes during an infection and is quite specific to the phage infecting the bacteria[206]. By being regulated by major signaling molecules or metabolism proteins, CRISPR-Cas systems could respond to such changes. In *Pectobacterium atrosepticum*, the type I-F CIRSPR-Cas system is regulated by the signalling molecule cAMP and its receptor CRP, as well as the metabolic enzyme GalM [204]. LeuO which can relieve H-NS repression is itself a transcriptional regulator that responds to amino acid starvation [206].

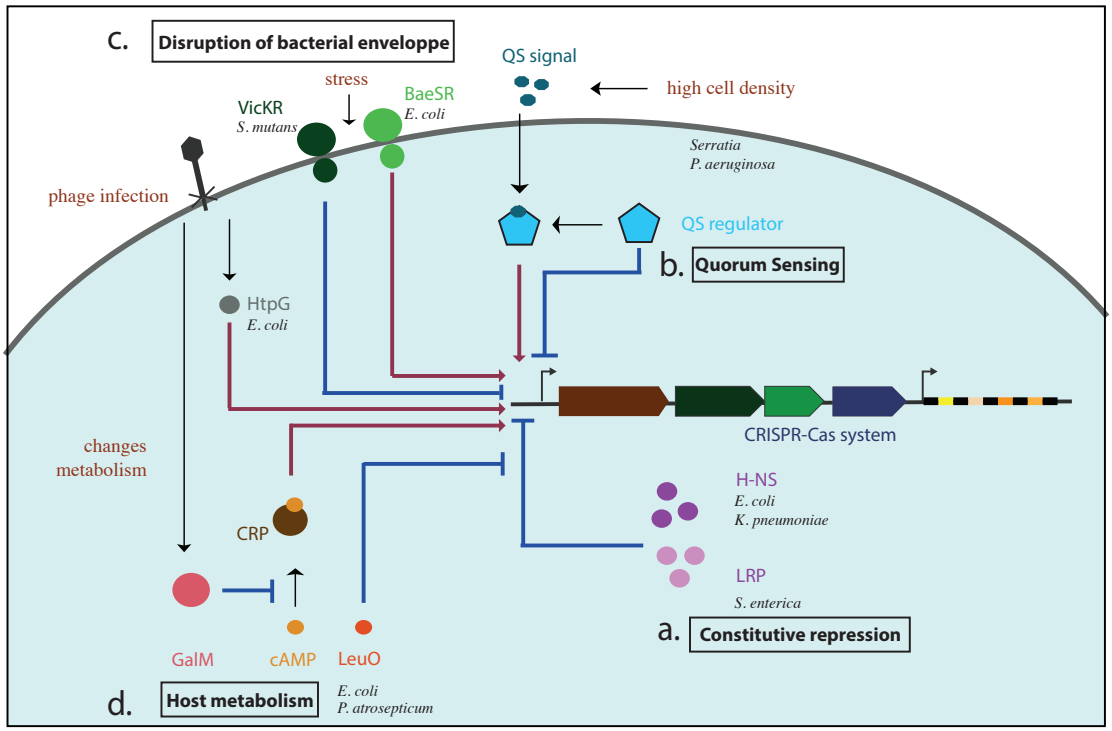


Figure 1.9: CRISPR-Cas systems regulation.
a. Constitutive repression. In bacteria like *E. coli*, *K. pneumoniae* and *S. enterica* CRISPR-Cas systems are repressed. **b.** Quorum sensing regulation. A high cell density will generate a quorum sensing signal (dark blue) which will alleviate the repression by a QS regulator (light blue). **c** Proteins associated to the bacterial envelope can regulate CRISPR-Cas systems. Both repression (VicKR system in *S. mutans*) and activation (BAeSR and HtpG) in *E. coli* have been reported. **d.** Metabolism regulation. Different metabolim molecules have been reported to activate CRISPR-Cas systems such as cAMP through its receptor CRP or LeuO by preventing the H-NS repression. Red arrows correspond to activation, blue lines to repression. Organisms in which regulation has been demonstrated are noted in black. Adapted from [206].

1.1.7 Where do CRISPR-Cas systems come from ?

According to a current hypothesis, CRISPR-Cas systems would have emerged through the insertion of a casposon next to a Cas10 like protein from which the effector module would have originated (Figure 1.10) [139].

Casposons are a class of self synthesizing transposons that encode a Cas1 homolog [142]. Comparative genomics of 62 strains of the archaeon *Methanosarcina mazei* allowed to show that these elements are mobile [143]. The integrase activity of the Cas1 homolog as well as the similar target site specificities of Casposon integration and CRISPR spacer incorporation have recently been demonstrated experimentally [112, 22]. The casposon might also have provided ancestors of

CRISPR repeats and the leader sequence [141]. Cas2 is thought to have originated either from the casposon or from a toxin-antitoxin module thus completing the adaptation module [140].

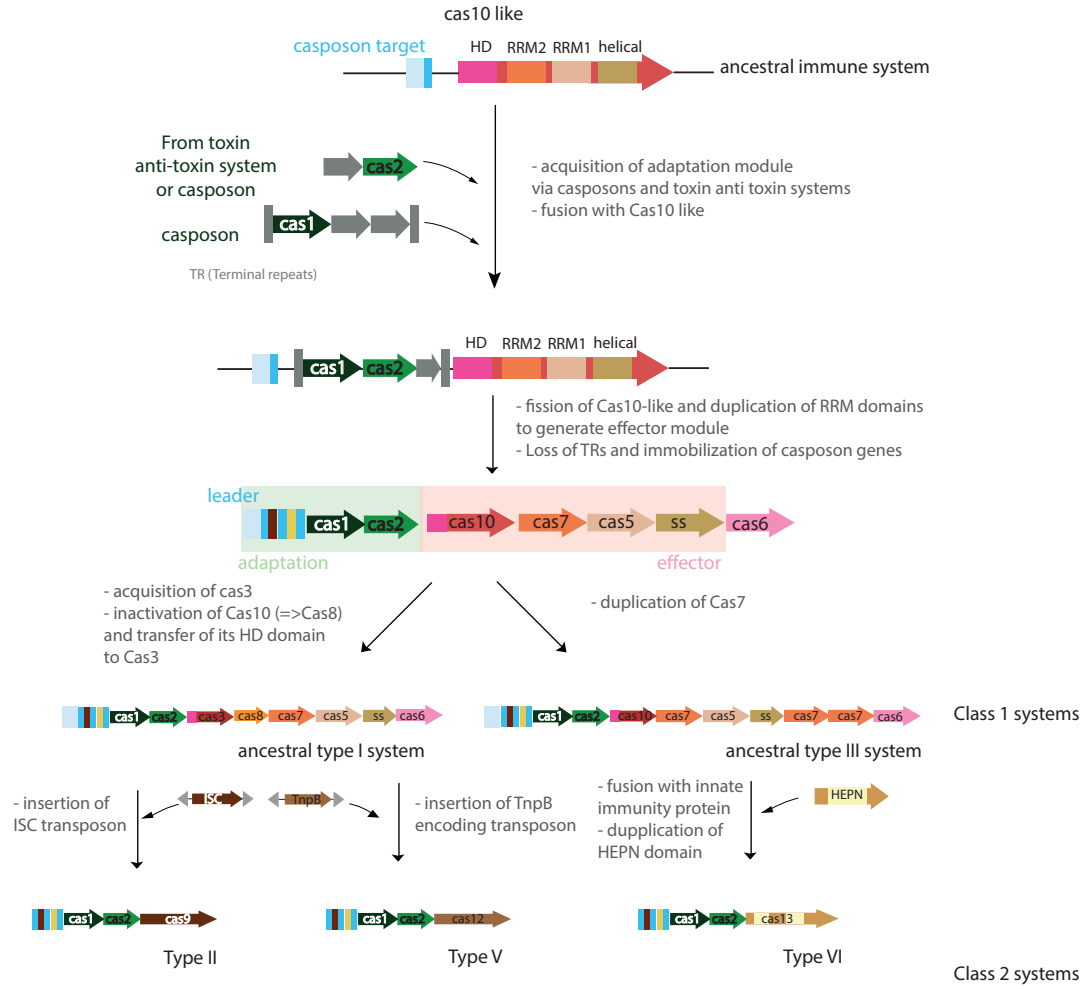


Figure 1.10: A working scenario for CRISPR-Cas systems evolution.
CRISPR-Cas systems would come from the association of a Cas10 like gene and a Cas1 homolog from a casposon. Further steps would have lead to the ancestral form of CRISPR-Cas systems which ressemble Class 1 systems. The multi protein effector would then have been replaced by single proteins from different mobile elements leading to the emergence of Class 2 systems.
Adapted from [181, 140, 141].

The ancestry of the effector module is even less clear. As Class 1 systems are widespread [162], Class 1 effectors are believed to be the ancestral form. As core subunits of Class 1 effectors contain divergent version of RRM domains, the hypothesis is that an ancestral protein like Cas10 evolved by serial duplication fol-

lowed by diversification [139]. The working hypothesis concerning Class 2 variants is that they evolved through replacements of the effector locus by single proteins that came from different mobile genetic elements. Type II systems would derive from proteins from IscB transposons [135] while type V systems are linked to TnpB transposons [244]. Finally, type VI ancestors have not been determined precisely yet, but would derive from Cas proteins containing HEPN domains [140].

In less than a decade, incredible progress has been made to understand how CRISPR-Cas systems work. However, not every aspects of CRISPR-Cas systems were tackled at the same speed. While initial reports focused on the immunity phase, the last few years, more studies have been dedicated to adaptation. This chronological aspect can be appreciated through the precise molecular details produced on different subtypes for the immunity phase which does not have a match for the adaptation phase. Even if many studies tackling molecular mechanisms were motivated by potential applications, the discovery of an adaptive immune system in bacteria also generated a lot of enthusiasm in another community, scientist studying the arms-race between bacteria and their phages and the subsequent consequences on bacterial evolution.

1.2 CRISPR-Cas systems impact on bacteria and phages

As an immune system, CRISPR-Cas systems have a very strong evolutionary impact at a short time scale by for example allowing a bacteria to survive a phage infection. However, they can also influence long-term microbial evolution. First, they can limit horizontal gene transfer. Second, non-canonical functions of CRISPR-Cas systems have emerged and impact various cellular processes like gene regulation, virulence or dormancy. Third, the constant arms-race between bacteria and phages has led to the emergence of original mechanisms to evade CRISPR-Cas immunity. Even if CRISPR-Cas systems also target plasmids, very few studies have experimentally examined the interplay between CRISPR-Cas systems and plasmids, which is why most of the concepts explained and work cited concern phages.

1.2.1 The impact of CRISPR-Cas systems on bacterial genome

The first potential influence of CRISPR-Cas systems on bacterial evolution is its impact on horizontal gene transfer. Experimental work have demonstrated that CRISPR-Cas systems constitute a barrier to natural transformation [31, 316], conjugation [169, 155], and transduction through immunity against phages [19]. Limiting horizontal gene transfer can impact bacteria in many ways. For example, a study showed that the size of the genome for specific strains of *P. aeruginosa* was directly impacted by the presence of CRISPR-Cas systems [23]. Genomes harbouring active CRISPR-cas systems were on average 300 kb smaller than those without CRISPR-Cas systems or with inactive ones [23].

Apart from genome size, CRISPR-Cas systems have been found to impact the acquisition of pathogenic traits in *E. coli*. [86]. A study detected a negative correlation between the number of type I-E CRISPR repeats dataset and the presence of pathogenicity traits [86]. Similarly, a strong inverse correlation between the presence of a CRISPR-Cas locus and acquired antibiotic resistance was detected in *Enterococci faecalis* [202]. Preventing HGT can lead to host specialization. In *Legionella pneumophila*, CRISPR-Cas systems prevent the transfer of an episome that leads to an improved fitness in *Acanthamoeba* but a reduced ability to replicate in other hosts and conditions [221]. Finally another argument for a negative impact of CRISPR-Cas systems on HGT is their absence from species where HGT is important and their presence in closely related bacteria. For example, *Streptococcus pneumoniae* does not have a CRISPR-Cas system while most closely related *Streptococci* do [103]. It was also demonstrated that loss of competence for natural transformation was often followed by CRISPR loss underlying the evolutionary link between natural transformation and CRISPR-Cas immunity [133].

However other studies have brought up evidence showing that CRISPR-Cas systems do not impact negatively HGT rate. Bacteria which encode at least one prophage (phage integrated in the bacterial chromosome) were more likely to harbour type I and type II CRISPR-Cas systems [269]. In *E. coli*, no specific link could be made between strains lifestyle and the presence or content of CRISPR-Cas systems [270]. Still in *E. coli*, CRISPR-Cas systems had no impact on the spread of antibiotic resistance plasmids [271]. More generally, when horizontal gene transfer was studied at evolutionary timescales by using three different measures [94], no correlation was found between the number of HGT events and the length of the CRISPR array suggesting that CRISPR-Cas systems do not impact HGT at this scale [94].

Overall, the impact of CRISPR-Cas systems on HGT seems to depend on the scale and the organisms in which studies were conducted. Even if experimental evidence and bioinformatics correlations in specific species tend to show that HGT is impacted by CRISPR-Cas systems, this does not hold for other organisms or at a larger scale. The impact of CRISPR-Cas systems on horizontal gene transfer and therefore on bacterial evolution remains a complex and open issue.

1.2.2 Non canonical functions of CRISPR-Cas systems

Apart from its canonical role as an immune system, several studies have shed light on other functions of CRISPR-Cas systems. A report has shown that some CRISPR arrays do not present the evolutionary dynamics of an active immune system, as they evolve much slower, but they are still conserved indicating other potential functions [272]. Several examples of these have now been demonstrated in diverse aspects of bacteria life including population behavior control, host-pathogen interactions and cell dormancy [293, 17, 222].

Population behaviour control

Two main examples have implicated CRISPR-Cas systems in population behavior control: the regulation of fruiting body in *Myxococcus xhanthus* and biofilm regulation in *Pseudomonas aeruginosa* (Figure 1.11.a) [222]. *M. xanthus* harbours a type I-C and a type III-B CRISPR-Cas system. Disruption of *cas5*, *cas7* and *cas8* of the type I-C leads to reduced sporulation [280]. Even if the mechanism is still unclear, Cas8 is a regulator of FrutA, a fruiting body transcriptional activator, which itself is involved in the expression of the Cas cluster. The Cas operon is only transcribed in certain growth conditions and the *cas* genes expressed in specific locations of the fruiting body [280, 293]. The formation of fruiting body is also influenced by type III-B CRISPR-Cas system which regulates exopolysaccharide (EPS) production and type IV pili-mediated chemotaxis.

Another population behaviour control involving CRISPR-Cas systems concerns biofilm formation. In *Pseudomonas aeruginosa*, the presence of the prophage DMS3 and a type I-F CRISPR-Cas system inhibits biofilm formation and swarming motility [311]. The molecular mechanism require all the interference components [44]. More specifically, a spacer of the type I-F CRISPR array targets a gene from the prophage DMS3 [44].

Host pathogen interactions

Most of the described non-canonical functions of CRISPR-Cas systems are linked to host pathogen interactions *i.e.* the ability of bacteria to infect an host (Figure 1.11.b). However this might reflect more a bias in microbiology research than an actual trend specific of CRISPR-Cas biology. One of the best characterized example is the regulation of virulence of *Francisella novicida* by its type II-B CRISPR-Cas system. Mutations in the CRISPR-Cas system lead to reduced virulence in mice. [234, 233]. The CRISPR-Cas system represses the production of a protein called BLP which is recognized by the host immune system. By repressing its production, *F. novicida* escapes host detection and increases its virulence [234]. The repression involves Cas9, the tracrRNA and the crRNA [233]. Other Type II systems have been shown to regulate virulence. Mutants of Cas9 in *Campylobacter jejuni* which harbour a Type II-C system, showed reduced adherence, invasion, and attenuated cytotoxicity towards human gut cell lines [156].

Similarly, Cas proteins of the type I-F CRISPR-Cas system of *Pseudomonas aeruginosa* strain UCBPP-PA14 target the mRNA of LasR, a quorum sensing regulator. This regulation reduces pro-inflammatory responses of the host in cell and mouse models [228]. In *Xenorhabdus nematophila*, a mutant in a region called NilD abolished the colonization of *X. nematophila* host [278]. The NilD RNA is actually a single CRISPR RNA and its expression depends on the presence of the Cas6 protein from a type I-E CRISPR-Cas system but not on the nuclease Cas3 [278].

In *Legionella pneumophila*, mutants for *cas2* cannot survive within amoebae and introducing *cas2* in strains which lacked it increased infectivity [97, 98]. Cas2 is upregulated during intra-amoeba growth and even if the regulation mechanism is not fully understood, the RNase activity of Cas2 is required [98]. Cas2 was also introduced in *Mycobacterium smegmatis* and led to an altered expression of sigma factors which are involved in mycobacterial stress response and virulence [117]. Finally, CRISPR arrays alone can impact virulence. *Listeria monocytogenes* harbours an orphan CRISPR array called Rlib. When overexpressed, RliB upregulates feoAB, a ferrous iron transporter involved in virulence [166].

Dormancy

Another role for CRISPR-Cas systems had been previously hypothesized: its implication in cell dormancy [293]. As detailed in the section on molecular mechanisms of immunity, recent studies have unveiled two mechanisms by which CRISPR-Cas systems could lead to toxicity, dormancy or cell death: 1) the non specific RNAse activity of Cas13a (C2c2) from type VI CRISPR-Cas system [2, 69] 2) Csm6, which upon activation by coA cleaves RNA non specifically [136, 189] (Figure 1.11.c).

The roles of CRISPR-Cas systems beyond immunity are diverse and involve both Cas proteins and CRISPR arrays. Even if specific cases have been reported, much remains to be understood on the prevalence of those alternative functions of CRISPR-Cas systems and how those functions evolved.

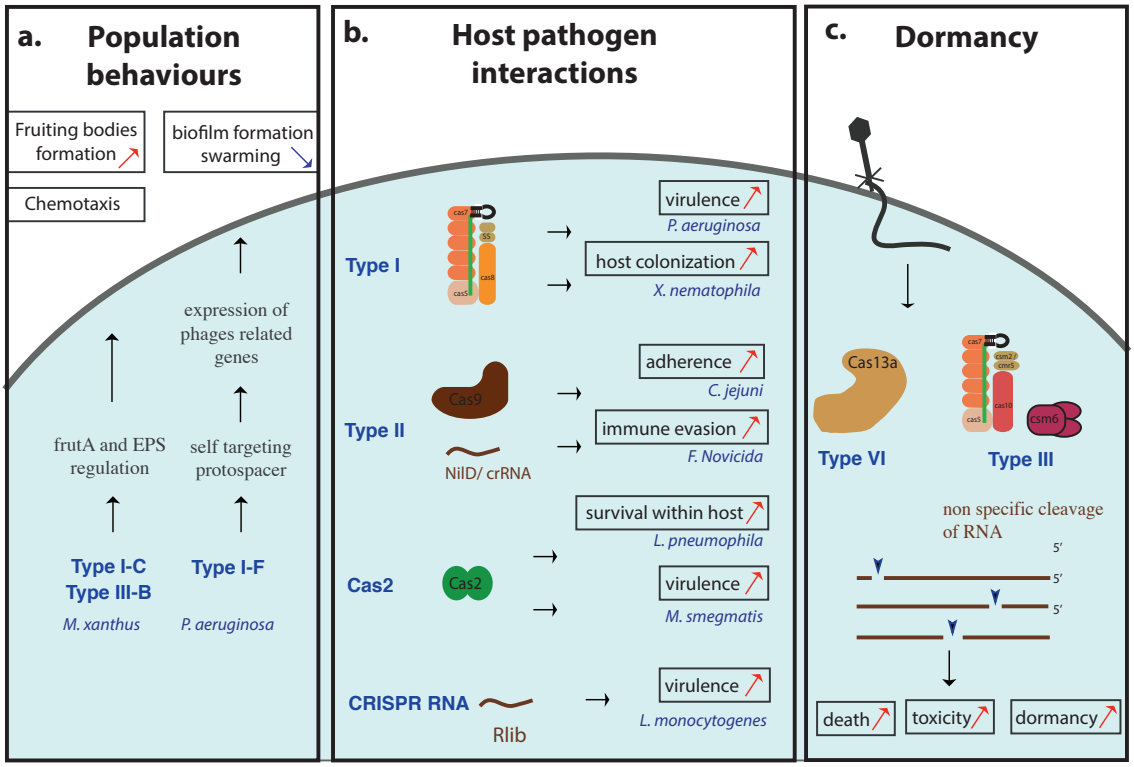


Figure 1.11: Non canonical functions of CRISPR-Cas systems.
Several non canonical function of CRISPR-Cas systems have been described in a wide range of bacteria and concern three main types of cellular processes: **a.** control of population behaviour, **b.** host pathogen interactions and **c.** dormancy. Organisms in which functions were described are indicated in blue and italic. Red arrows correspond to enhancement by CRISPR-Cas systems, blue arrows to decrease

1.2.3 CRISPR impact on phages: the consequences of a constant arms race

Bacteria and phages are engaged in a constant arms race. As with any defense system, phages have found ways to escape CRISPR-Cas immunity using a variety of mechanisms which involve genome modification or specific anti-CRISPR proteins (Figure 1.12).

Avoiding sequence specific targeting through genome modifications

Phages have a high mutation rate which enables them to adapt rapidly [67]. It was quickly hypothesized and then observed that phages could escape CRISPR-Cas systems by single mutations. Several co-evolution experiments between bacteria harbouring a functionnal CRISPR-Cas systems and a phage have been led in *Streptococcus thermophilus* [62, 263, 199] and in *Synechococcus* [106]. They report the accumulation of single mutations in phages in CRISPR targeted regions. More specifically, the PAM sequence is more frequently mutated compared to the protospacer. As in some CRISPR-Cas subtypes, there is a tolerance for sequence specificity and priming effects. Mutations in the PAM ensure complete abolition of CRISPR immunity (Figure 1.12.a).

Co-evolution experiments have also led to the observation of genome rearrangements [199]. It was observed that the presence of multiple phages increased phage persistence. Genomes analysis revealed that recombination events took place between phages thus generating chimeric phages which did not harbour CRISPR targeted regions [199]. Similar observations have been made when studying natural ecosystems. Extensive recombination events have been observed in phages of two natural acidophilic biofilms to escape CRISPR-Cas immunity (Figure 1.12.a) [9].

Phages can also modify their genomes through the attachment of chemical groups to their DNA [41]. These forms of DNA modifications are used to escape other bacterial immune systems like restriction modification systems [286]. Wild-type T4 containing glc-HMC, a covalent modification of the DNA where cytosine is replaced with glucosyl-hydroxymethylcytosine, was resistant to CRISPR-Cas9 interference [41]. However, genome modifications are not the only way by which the arms-race operate.

Anti- CRISPRs

Phages encode small proteins that inhibit CRISPR-Cas systems and were named anti-CRISPRs or Acr (Figure 1.12.b)[252, 37]. Discovery of anti-CRISPR came through the unexpected observation that lysogenized strains of *P. aeruginosa* with specific prophages were susceptible to lytic phages while the non lysogenized ones

were resistant [35]. This was explained by the discovery that the prophages encoded proteins inhibited the otherwise active Type I-F CRISPR-Cas systems. Acr have now been found against type I-E, I-F, II-A and II-C CRISPR Cas systems [208, 207, 223, 118, 37]. They are encoded by both temperate and lytic phages [118, 37]. They present different mechanisms of actions. They can prevent target recognition by binding to the effector complex or to the nuclease in type I systems [283, 285, 51]. Similarly, they can hamper sgRNA-guided DNA binding through DNA-mimicking in type II systems [65, 307, 241].

Anti-CRISPR are widespread. 64% of 449 *P. aeruginosa* type I-F CRISPR-Cas are inhibited by chromosomally encoded AcrF [209]. It was further estimated that 53% of 81 *P. aeruginosa* type I-E CRISPR-Cas are susceptible to AcrE[36] and 50% of *L. monocytogenes* type II-A systems to AcrIIA [223]. The co-occurrence of CRISPR-Cas systems and Acrs leads to interesting consequences for both phages and bacteria. If a phage harbours an Acr, this may allow lysogeny and thus HGT [37].

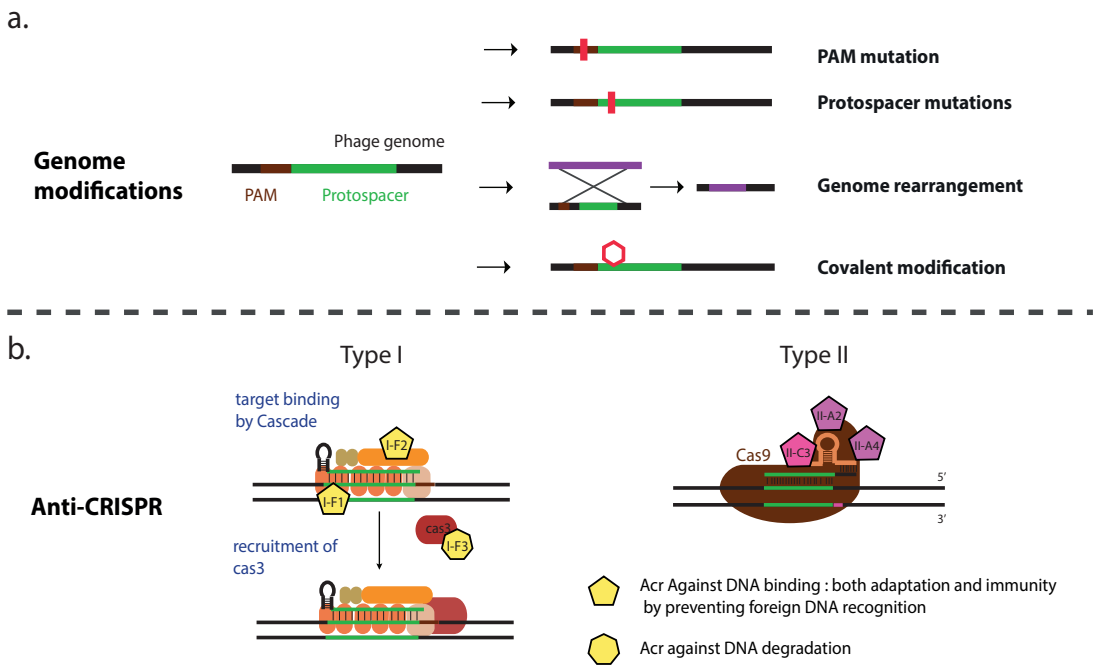


Figure 1.12: Phages escape mechanisms from CRISPR-Cas immunity.
a. Genome modifications such as mutations in the PAM sequence, or in the targeted sequences, genomes rearrangements or covalent modification have been shown to allow phages to escape CRISPR immunity. **b.** Anti-CRISPR proteins or Acr (yellow or pink polygones) inhbitit CRISPR-Cas systems in a type and even subtype dependent way, by binding to different domains of Cas proteins involved in target recognition or DNA cleavage.

The recently discovered anti-CRISPRs, beyond constituting an interesting manner of escaping CRISPR-Cas immunity generated a lot of interest from the CRISPR-based technology community. These small proteins could indeed be used to regulate proteins such as Cas9 which could prove very useful for specific applications.

The diverse mechanisms to escape CRISPR-Cas immunity introduced in this section directly underline one major challenge for CRISPR-Cas systems : how to deal with escape mutants. The arms-race between bacteria and phages obviously also impact the way CRISPR-Cas systems evolve. The main dynamics, and factors that influence this evolution constitute the focus of the next section.

1.3 Evolution of CRISPR-Cas systems

1.3.1 CRISPR-Cas systems are scattered

CRISPR-Cas systems are widespread in bacteria thanks to horizontal transfer

CRISPR-Cas systems are found in more than 90% of archaea and in around 50% of bacteria [161]. They are present across many bacteria phyla [161]. Many metagenomics studies have revealed the presence of CRISPR-Cas systems in very diverse environments including gut [225, 260, 91, 167], rumen [26], acid mine drainage [9, 273], antarctic surface snow [154] or hot springs [106, 107].

This widespread distribution can partly be explained by the observation that CRISPR-Cas systems are massively transferred horizontally. First evidence for this transfer was discovered in the early 2000s before CRISPR-Cas mechanisms were understood. Phylogenetic analysis of *cas* genes and presence in diverse prokaryotes led to suggest that HGT was responsible for the movement of these clusters among distantly related genomes [159]. The first study fully dedicated to the HGT of CRISPR-Cas systems was performed in 2006. The phylogenetic analysis of *cas* genes revealed incongruences with the species tree indicating HGT [90]. Further analysis confirmed the high level of transfer either using comparative network clustering of direct repeats (DRs) and *cas* genes [47] or focusing on archaea [11] or on specific genera like *Shigella* [306], *Campylobacter* [210].

This horizontal transfer is also supported by the fact that CRISPR-Cas systems have been detected on diverse mobile genetic elements (MGE). CRISPR-Cas systems were found on megaplasmids in diverse bacteria including *Treponema denticola*, *Vibrio vulnificus*, *Yersinia pestis* [90]. Functionning type I and III CRISPR-Cas systems were characterized in a plasmid of the *Cyanobacterium synechocystis* sp. PCC6803 [235], in a *Lactococcus lactis* [177] and on *Streptomyces* linear plasmid pSHK1 [99]. Apart from plasmids, phages also encode CRISPR-Cas systems. First, CRISPR arrays were detected in *Clostridium* prophages [253]. Metagenomic studies of the human gut virome revealed phages encoded CRISPR [178]. Experimental evidence came with the exemple of the Vibrio cholera phage ICP1 which encodes a functional Type I-F CRISPRCas system and uses it in an ironic turn of events to neutralize phage resistance mechanisms [236]. Finally, recent evidence demonstrated that Tn7-like transposons encode type I-F and type I-B minimal CRISPR-Cas systems [212].

CRISPR-Cas systems remain quite rare for an immune system

Despite important transfer, CRISPR-Cas systems remain rare in bacteria. While several studies on fully sequenced genomes have reported their presence in 90% of

archaea and 50% of bacteria [161, 1], a recent report on groundwater filtrates revealed that in this environment only 10% of archea and bacteria encoded CRISPR-Cas systems [43]. More strikingly, some phyla were completely devoid of CRISPR-Cas systems [42]. Study of these phyla in other environments confirmed that the absence of CRISPR-Cas systems was not restricted to a specific environment [42]. The gap between those detections can be explained by the genomes studied. The abundance of certain clades in groundwater filtrates differ from their abundance in the database of fully sequenced genome and might suggest that CRISPR-Cas systems are less advantageous in this specific environment. Altogether, their prevalence appear quite low in comparison to other defense systems such as restriction modification, of which two copies per genome are found on average [197]. Even in the same species, strains differ in terms of presence of CRISPR-Cas systems [294]. This scarcity raises questions relative to potential costs of harbouring CRISPR-Cas systems.

Several studies have also reported that CRISPR-Cas systems are not evenly distributed across these environments. More specifically, high-temperature environments encompass more organisms encoding CRISPR-Cas systems [8, 160, 289]. An enrichment in CRISPR-Cas systems was also reported in marine sponge-associated microbial metagenomes [114] while groundwater filtrates showed reduced prevalence of CRISPR-Cas systems [42]. This scattered distribution underlines the impact of ecological factors on CRISPR-Cas systems distribution.

CRISPR-Cas systems are lost and gained

To explain the variation in the presence, number and function of CRISPR-Cas loci within and between bacterial species, several studies argue for a dynamic of frequent loss of CRISPR-Cas systems that are then gained again through horizontal gene transfer[266, 25, 128]. *Cas* genes evolve under purifying selection that is typically much weaker than the median strength of purifying selection which suggest that they could be lost easily [266].

An experimental study in *Staphylococcus epidermidis* led to the observation that the loss of CRISPR-Cas systems even by large deletions can have little or no fitness cost *in vitro* [128]. Loss of CRISPR-Cas systems was observed in different set ups. In *Thermoanaerobacter* genus, transposons insertions in CRISPR arrays were detected in six out of 8 species carrying CRISPRs [157]. In *S. agalactiae*, removal of spacers was observed repeatidly. In *Mycoplasma gallisepticum*, a host shift was followed by a very fast inactivation and then loss of CRISPR [59]. The evidence for frequent loss and frequent transfer led to the assumption that CRISPR-Cas systems are in a continuous state of flux by being lost when they bear a cost or when they are neutral and then reacquired by horizontal transfer (Figure 1.13)[128].

1.3.2 Evolutionary dynamics of CRISPR arrays

Dynamic and diverse CRISPR spacer content targets the mobilome

The dynamics of CRISPR arrays spacer content were examined in diverse environments. Observations on acidophilic biofilms revealed rapid acquisition of new spacers in different CRISPR loci [9, 273]. The study of the saliva from four human subjects over an 11 to 17 months time period revealed highly dynamic and diverse CRISPR arrays. A core (ranging from 7% to 22%) of shared CRISPR spacers remained stable over time within each subject, but nearly a third of CRISPR spacers varied between time points. Even in Antarctic surface snow, thousands of unique spacers were recovered with less than 35% overlap between four sampling sites [154]. In a study of biofilms in acid mine drainage, despite the recovery of 452 686 spacers in the bacterial population, rarefaction curves of spacers showed no saturation [264]. Corroborating metagenomics observations, many experimental studies show rapid acquisition (less than 24 hours) of diverse spacers in different species including *Sulfolobus solfataricus*, *S. thermophilus*, *E. coli*, *P. aeruginosa* [77, 263, 149, 199, 145, 277]. However, these experiments may not have been representative of natural conditions as they were achieved with one phage and one strain. The diversity of spacers observed in CRISPR arrays has even been used for pathogen tracking using spoligotyping. It was employed in diverse Gram-positive and Gram-negative bacteria, including *Mycobacterium tuberculosis*, *Salmonella enterica*, *Yersinia pestis* and *Erwinia amylovora* [21, 240].

The rates of acquisition and loss of spacers are not the same along the CRISPR array. In metagenomics samples of acid mine drainage biofilms, the spacers found at the trailer end (opposite to the leader end) and therefore the oldest are conserved for at least 5 years [264]. Similar observations were made in metagenomics analysis of the gut microbiome [225, 260]. Blocks of historical trailer-end spacers were shared by multiple individuals [176]. Thus, CRISPR array dynamics can be described as leader end diversity and trailer end clonality, meaning that in a population, the leader end (near the transcription start) will encode very diverse spacers while the trailer end of the arrays will harbour similar spacers (Figure 1.13).

However, not all CRISPR-Cas systems have a high rate of acquisition of new spacers. CRISPR array evolution in *Escherichia coli* and *Streptococcus agalactiae* can be explained mainly by vertical inheritance and differential spacer deletion [145]. *E. coli* strains which diverged around 250 thousand years ago show identical CRISPR [272]. These dynamics led to question the role of CRISPR-Cas as an adaptive immune system in this species.

Despite its role in immunity, several studies have reported that spacer sequences from metagenomic data rarely match viral genomes [8, 91]. Several hypothesis have

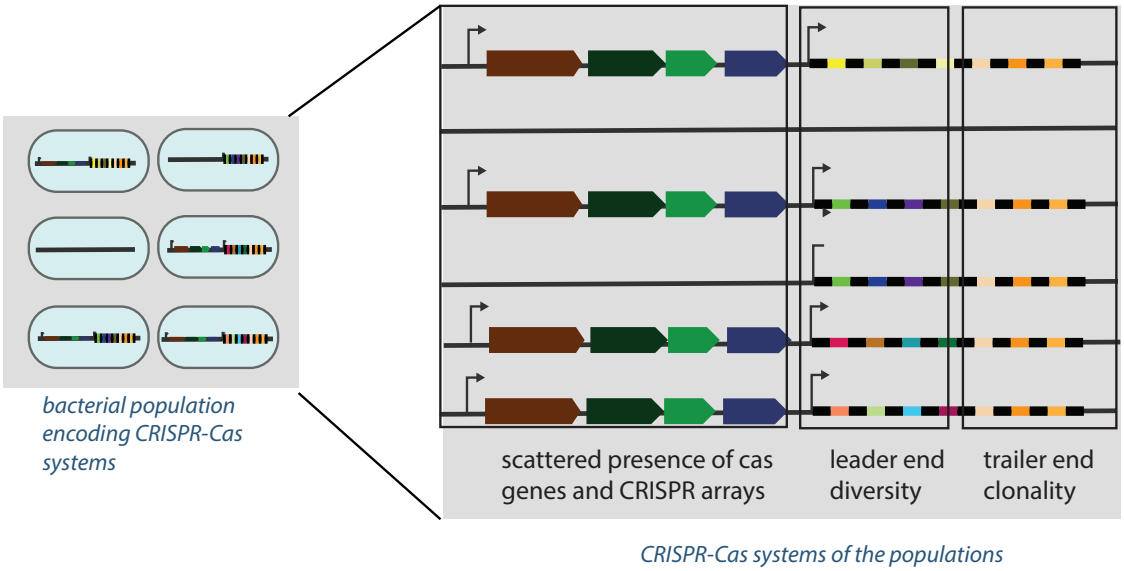


Figure 1.13: The evolutionary dynamics of CRISPR-Cas systems in a bacterial population.
CRISPR-Cas systems are often lost and then gained by horizontal gene transfer which leads to a scattered presence of *cas* genes and CRISPR arrays. The CRISPR arrays at a population level are characterized by leader end diversity (diverse spacers near the promoter) and trailer end clonality (similar spacers at the end of the array).

been put forward to explain this paradox. First, few viral sequenced genomes are available. Findings that spacers match phages present in the same sample corroborate this hypothesis [9, 260]. Second, given the fast evolution of phages to escape CRISPR-Cas immunity, spacers can match sequences that no longer exist, as phages bearing these sequences disappeared. This is corroborated by the fact that only recently acquired spacers match coexisting phages [9, 260]. A third hypothesis is that CRISPR would only target rare phages, making them more difficult to sequence [75]. In a metagenomics study, the vast majority of spacers did not match any of the 140 viral contigs $>$ 10 kb, which came from virus abundant enough to be assembled easily [75]. However, no experimental studies have corroborated these observations yet. Recently, a quantitative and comprehensive investigation on CRISPR spacers and their targets from fully sequenced and draft bacterial and archaeal genomes was performed [245]. Targets could be identified for 7% of the spacers and originated almost exclusively from MGE DNA. Oligonucleotide composition comparison between spacers and MGE DNA led the authors to conclude that the remaining 93% also originates from the mobilome [245].

Hypothesis to explain the observed dynamics

Several models have been proposed to explain the observed trailer-end clonality and leader-end diversity of CRISPR arrays.

The different models are built on three basic theoretical frameworks of host-pathogen co-evolution [76]. First, diversity could be generated by trade-offs between immunity and fitness. While encoding more CRISPR-Cas systems with diverse spacers could increase immunity, it might reduce fitness when phages infections are low as resources are invested in a useless defense system. Such models predict that maintenance of CRISPR-Cas systems if viral diversity is limited [102]. Second, diversity could be maintained through negative frequency-dependent selection, where the adaptive benefit of a novel allele decreases as it increases in frequency within a population. If a spacer is frequent in a population, it is more likely that escape phages will arise than for a rare spacer. Thus, a bacteria encoding a specific spacer might become less fit if this spacer is shared by others in the population. This would lead to the maintenance of spacer diversity. Third spatial structure could play a role in CRISPR diversity [101]. A model predicts that self organization of bacteria in space would lead to the co-existence of subpopulations of bacteria with a diversity of spacer numbers, with intermediate spacer numbers most frequent [101].

Taking these hypothesis into account, more complex models investigated the evolutionary dynamics of CRISPR arrays. One suggested that neutral variation persists at the leader end until selection for a particular spacer causes a selective sweep [105]. In certain conditions, among spacers that were equally fit, one spacer or a set of spacers will become more advantageous thus invading the population. New diverse spacers will then be integrated at the leader end thus generating arrays with similar spacers at the trailer end and diverse ones at the leader end. A second approach led to similar predictions that the selection for trailer end clonality is caused by rapide selective sweeps by highly immune CRISPR lineages [288]. In that framework, the leader end conservation is explained by the selective advantage of preexistng spacers against persisting viral sequences [288].

A third approach suggested that co-existence of diverse spacers targeting the same phage could be considered as equally-fit immune alleles and would confer distributed immunity to a microbial population [50, 49]. The key factors for the emergence of the co-evolutionary model is the asymmetry between 1) the vast reservoir of protospacers due to the limited cost of generating a new allele 2) the fitness constraints of evolving escape mutations for phages enhanced by the fact that an escape mutant will only be resistant to one allele [49]. Using the concept of distributed immunity, a model could predict a sustained diversity and stability of CRISPR-Cas systems. This stability could lead to a decrease in viral population density and even to viral extinction. The analysis of available experimental data of

coevolving populations of *Streptococcus thermophilus* and their viruses showed the rapid emergence of distributed immunity in the host population [49]. Distributed immunity is also influenced by viral mutation rate, and host acquisition rate [49].

The concept was further validated by experimental evidence. In *in vitro* co-evolution experiments between *P. aeruginosa* and *S. thermophilus* and their respective viruses, spacer diversity led to viral extinctions as viruses could no longer escape CRISPR-Cas immunity by point mutation [277]. This was confirmed by experiments demonstrating that phages could become locally adapted to their bacterial hosts but only when CRISPR allele diversity was low [184].

However underlying hypothesis of the distributed immunity model might need to be adjusted. A strong bias in spacer selection was detected during co-evolution experiments between *S. thermophilus* and a bacteriophage as a few spacers were much more abundant than what could be expected from a random selection. This suggests either that some spacers are better than others or that there is a limit towards the number of spacers that can be acquired as some are more preferentially captured [198].

1.3.3 Ecological factors and the relative benefits of CRISPR-Cas systems

Bacteria possess an arsenal of diverse immune systems to counteract phages. Given the relative abundance of CRISPR-Cas systems compared to other defense mechanisms, understanding how those different defense systems interact and the conditions in which specific systems will be favored over others is essential. A study in *S. thermophilus* demonstrated that CRISPR-Cas and restriction-modification systems could be compatible and even work in a synergistic manner resulting in increased phage resistance [68]. However, in another experimental set up, *P. aeruginosa* either develops CRISPR-Cas based resistance or surface modification to overcome phage DMS3 showing that defense systems don't always function synergistically [296].

A key factor for the emergence of CRISPR-Cas systems is low phage diversity [289, 120]. Two different models predicted that benefits of CRISPR-Cas systems would decrease as virus genetic diversity increases [289, 120]. Above a threshold value of total viral diversity, the CRISPR-Cas system would become ineffective as it would not be able to encode enough spacers to target the diverse phage population and would therefore be lost. As genetic diversity depends among other determinants on mutation rates and population size, both of these factors will impact the maintenance of CRISPR-Cas systems. In one of the studies, the authors argue that as mutation rates are typically higher in mesophilic prokaryotes than in thermophilic prokaryotes, the impact on viral diversity would explain the higher

presence of CRISPR-Cas systems in thermophiles [289]. Another approach reached similar conclusions by showing that CRISPR-Cas like systems would be favored against slowly evolving bacteria [172]. The higher prevalence in thermophiles could also be explained through variation of the second parameter, population size [120]. As populations of mesophiles are larger than thermophiles, higher phage diversity would occur in a large population. CRISPR-Cas systems would become ineffective in that context [120]. Finally, viral diversity can also allow new escape mechanisms such as recombination between phages susceptible to a specific spacer and immune ones, limiting even more CRISPR-Cas immunity [199].

A second key factor for the emergence of CRISPR-Cas systems is low phage abundance. Key findings on the importance of the force of infection, that is to say the number of infective phages, come from an experimental study in *P. aeruginosa* [296]. In the experimental setup, *P. aeruginosa* can evolve either surface modification or CRISPR-Cas system to escape DMS3 phage. When performed in a high resources media such as LB, the co-evolution experiment led to the emergence of almost exclusively surface modification while in a low ressource media, it was the opposite with almost exclusively CRISPR-Cas resistance. These observations were explained by one parameter: phage abundance. When faced with rare infections, an inducible system like CRISPR-Cas will be favoured, while constant infections will favour a constitutive type of defense. Modeling of diverse immune strategies reached the same conclusions [172]. Phage abundance can be impacted by other factors. For example, migration of phage-sensitive host in a population harbouring mainly CRISPR-Cas system can lead to an increase in phage abundance. In a similar experimental set up using *P. aeruginosa*, it was shown that such migration led to a switch in defense mechanisms in the population : from CRISPR-Cas systems to surface modification [115].

1.3.4 Costs associated to encoding a CRISPR-Cas system

The relative scarcity and frequent loss of CRISPR-Cas systems suggest that they also bear costs including: costs of maintaining and expressing a CRISPR-Cas system, costs linked to autoimmunity phenomenon and costs from limiting horizontal gene transfer (Figure 1.14).

Maintenance and expression of CRISPR-Cas systems

It is quite difficult to pinpoint the cost of expressing a CRISPR-Cas system as CRISPR-Cas systems are very diverse in protein content and regulated in different manners. The cost of Cas proteins expression and of acquiring new spacers were measured in *S. thermophilus*. The main cost came from the expression of the Cas proteins while additional spacers were not associated with any costs. This led the authors to conclude that the cost of the CRISPR-Cas system was mainly due to the maintenance of the defense system [274]. In a second study on *P. aeruginosa*,

CRISPR-Cas systems were found to have an infection dependent fitness cost but no constitutive cost [296]. As both studies reach different conclusions, further studies on different species and types of CRISPR-Cas system could help clarify this issue.

Autoimmunity

A second potential fitness cost generated by CRISPR-Cas systems is linked to autoimmunity. Self targeting spacers were detected in the first studies that postulated the immune function of CRISPR-Cas systems [217]. In *Yersinia pestis*, out of 36 spacers analyzed, 8 spacers matched sequences on the *Yersinia* chromosome [217]. A more general study of CRISPR arrays from 330 organisms found that one in every 250 spacers is self-targeting, and that such self-targeting occurs in 18% of all CRISPR-bearing organisms [259]. Experimental evidence in *E. coli* showed that in this experimental model, the incorporation rate of self targeting spacer is around 1 to 1000 [149, 63]. Acquisition of self-targeting spacers have consistently been observed in various organisms in *in vivo* adaptation experiments demonstrating that this is not a rare event [111].

The most common outcome of this self targeting is cell death. It was first demonstrated in *E. coli* by showing that the lambda prophage was not protected from the CRISPR-Cas system [72]. Further studies consistentely led to similar observations in diverse species with diverse CRISPR-Cas systems [111].

Even if mortality due to autoimmunity is high, bacteria can survive through several means. The pressure of self targeting can lead to genome remodeling. In *Pectobacterium atrosepticum*, the type I-F CRISPR-Cas system has a spacer targeting its own chromosome, more specifically a plant pathogenicity island. The strong toxicity generated by self targeting imposes a strong selective pressure that can result in large-scale genomic alterations, including the remodelling or deletion of entire pre-existing pathogenicity islands [279]. Similar results were shown with the type II-A CRISPR-Cas system of *S. thermophilus*. When confronted with self targeting spacers, 99% of the bacteria died but the surviving transformants contained large deletions of the targeted regions which were the result of recombination between insertion sequence elements (IS) [237].

Limitation of horizontal gene transfer

Providing resistance towards foreign genetic elements through for example CRISPR-Cas immunity can in some cases be deleterious. Indeed, bacteria adapt through horizontal gene transfer, and limiting the flux of HGT could prevent the bacteria from acquiring desirable traits. Two studies have brought clues to support this claim. First, a study looked at the relative advantage of encoding a CRISPR-Cas system or acquiring an antibiotic resistance gene through plasmid transfer in *S.*

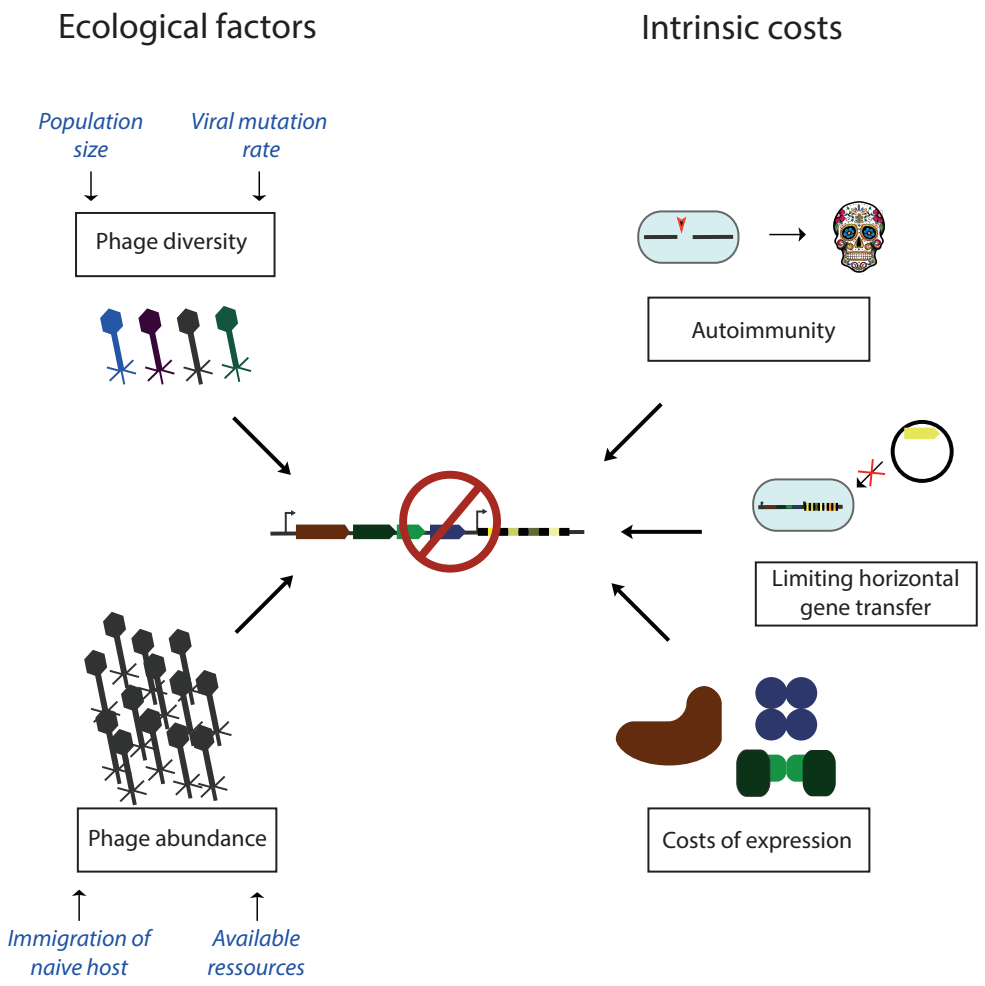


Figure 1.14: The downsides of CRISPR-Cas systems.
CRISPR-Cas systems relative scarcity can be explained by both ecological factors including phages diversity and phages abundance and intrinsic costs including cost of expression, limiting HGT and autoimmunity.

epidermis [128]. They programmed the CRISPR-Cas system to target the plasmid bearing the antibiotic resistant cassette. They found that when exposed to antibiotics, CRISPR-Cas mutants quickly emerged in this population thus favoring the acquisition of the plasmid with the antibiotic resistant cassette over the maintenance of the CRISPR-Cas system [128]. Similarly, another study repeated the seminal experiment that led to the discovery of horizontal gene transfer : *S. pneumoniae* without capsule genes were used to infect mice. Without capsules, infection fails [31]. However, when these strains are used to infect mice in the presence of heat-killed encapsulated *pneumococci*, infection succeeds thanks to the transfer of capsule genes from the heat-killed *pneumococci* to the alive acapsulated *S. pneumoniae*. When repeated with strains encoding CRISPR-Cas systems

targeting capsules genes, only mutants that inactivated the CRISPR-Cas system survived the mice immune system and achieved a successful infection. This suggests that the loss the CRISPR-Cas system was adaptive under these conditions [31]. Consistently with these results, an analysis of *C. jejuni* showed that strains with defective CRISPR-Cas systems carried a prophage or a virulence-conferring plasmid that were absent in two closely related strains with functional CRISPR-Cas systems. These observations underline the fact that the need for HGT may select against CRISPR-Cas systems.

Some mechanisms limit the different costs of CRISPR-Cas systems described above. First the regulation and more specifically the repression or inducible expression of CRISPR-Cas systems can reduce the intrinsic cost, autoimmunity effects and allow horizontal gene transfer. Similarly, immune priming has been proposed to limit autoimmunity effects and to select which MGE are targeted. This could allow for some differenciation between potentially beneficial or detrimental MGE [276]. At a short time scale, these mechanisms limit the costs and could favor the maintenance of CRISPR-Cas systems. In organisms, where CRISPR-Cas systems are regulated, their loss could therefore be explained by long period of absence of selection more than by the constitutive costs.

To conclude, CRISPR-Cas systems remain relatively rare for an immune system. While ecological factors seem to play a major role in their uneven presence in diverse environments, they might not explain their scattered distribution in bacterial genomes. Several hypothesis have been put forward to explain this paradox: costs of maintenance, of autoimmunity and of the limitation of horizontal gene transfer. However, none of these explanations fully solves this question. First, there is not clear reason why the cost of autoimmunity should vary between clades. Second, the cost of maintaining defense systems and preventing HGT are general for all defense systems. Thus, we propose a complementary explanation : the success of CRISPR-Cas acquisition by horizontal gene transfer is partly determined by the interactions of these systems with the genetic background of the host. To test this hypothesis, we decided to focus on one essential bacterial function: DNA repair. Our choice was motivated by several reasons. First, CRISPR-Cas systems and DNA repair pathways share the same substrate, DNA. Second, by potentially being able to repair breaks generated by CRISPR-Cas systems, DNA repair pathways could limit CRISPR-Cas efficiency.

1.4 DNA repair pathways in bacteria

1.4.1 DNA repair in bacteria : an overview

Several types of damages can affect DNA such as mismatches, single strand breaks or double strand breaks. They are repaired by a wide variety of pathways: base excision repair; mismatch excision repair; nucleotide excision repair; homologous recombination (HR); non homologous end joining (NHEJ) As CRISPR-Cas systems generate DNA breaks, we will focus on DNA repair pathways that mend such damages.

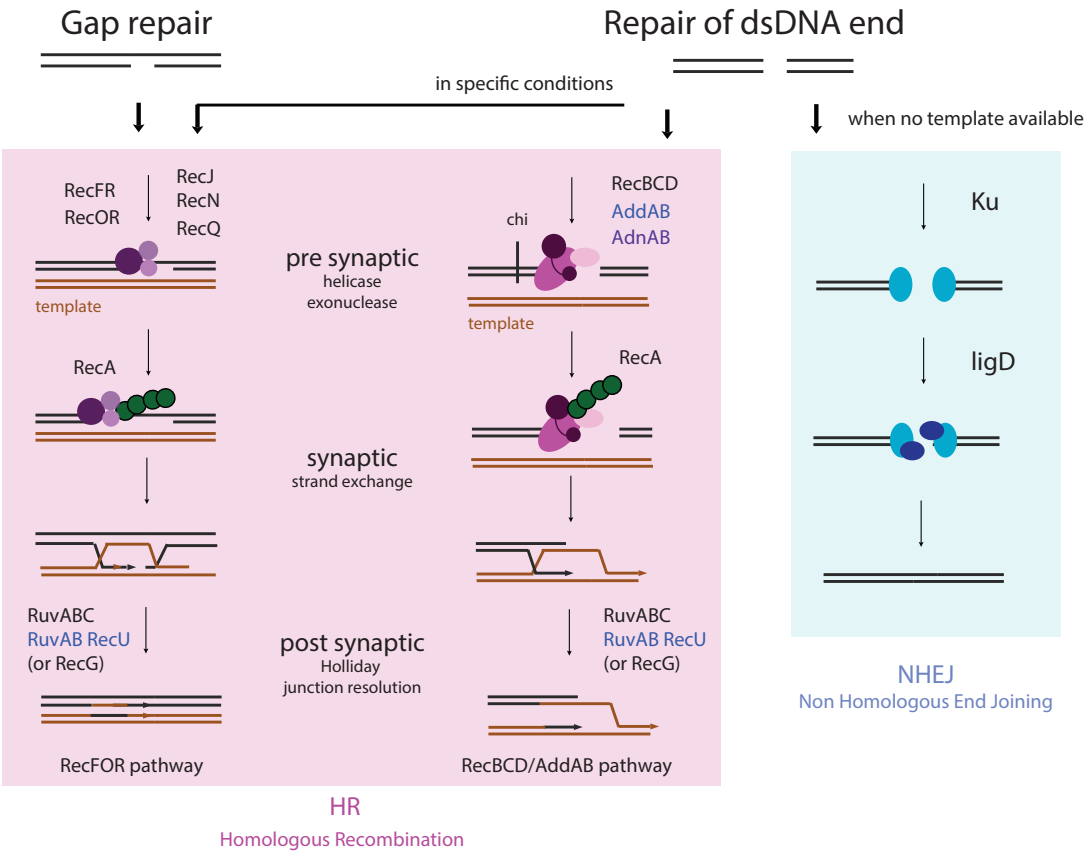


Figure 1.15: Simplified overview of DNA breaks repair in bacteria. Bacteria repair DNA breaks using mainly two pathways: homologous recombination (HR) and Non Homologous End Joining (NHEJ). Homologous recombination is organized in three phases: pre-synaptic, synaptic and post-synaptic. Pre-synaptic proteins recognize and process the DNA breaks and then recruit RecA. RecA catalyzes the synaptic phase which involves strand exchange. Post synaptic proteins resolve the Holliday junction. Mainly two pathways carry out HR : RecBCD/AddAB which repairs double strand breaks and RecFOR which repair mainly gaps even if it can mend DSB in specific contexts. NHEJ repairs double strand breaks without any template. Ku binds DNA ends and recruits ligD or other ligases to seal the break.

Bacteria repair DNA breaks using mainly two pathways: HR and NHEJ (Figure 1.15). HR can repair single strand gaps and double strand breaks while NHEJ only repairs double strand breaks (DSB) [175, 15, 4]. Homologous recombination requires an intact template while NHEJ does not require a template but is error prone [38]. Other systems called alternative end joining exist in some bacteria. They can repair DSB by ligation using microhomologies but they do so at very low frequencies [48, 100]. HR is by far the most common system. It is almost ubiquitous in bacteria while NHEJ is present in less than 30% of bacteria [227]. NHEJ is mainly used when HR is impaired or when no template is available [38, 171].

DNA repair is a very complex process. Bacteria can have proteins with overlapping functions and can use different pathways to mend the same type of breaks. More specifically, some pathways can be used as back up systems when other are impaired. Moreover, a lot of different regulators modulate different parts of the pathways [175, 4]. This complexity is reflected in the number of proteins involved in theses processes. Genetic screenings in *B. subtilis* mutants have allowed to identify at least 40 proteins involved in DSB repair [15].

1.4.2 Molecular mechanisms of homologous recombination

Most of the knowledge described here relies on the model organism *E .coli*, yet some differences have been found in other bacteria and are indicated throughout the section.

Pre-synaptic phase

The first step to repair a break is to sense and recognize the damage. In *B. subtilis*, RecN acts as a first responder to DSBs. It senses the damage and then recruits multiple DNA repair proteins. Bound to DSBs, it tethers DNA ends and works as a scaffold for other DNA repair proteins [15]. In *E. coli*, as its concentration is highly regulated and maintained at very low levels, except during the SOS response, it might play the same role of DNA tethering but only in very specific conditions [188]. The main actors of the pre-synaptic phase (RecBCD, RecFOR, AddAB and AdnAB) are multisubunit complexes that use different combinations of helicases and nucleases to carry out strand resection to generate a substrate for RecA. [15].

RecBCD is the main complex for repair of DSBs in *E. coli*. It is composed of two ATP-dependent helicase motors : RecB going 3' to 5' and RecD moving from 5' to 3'. Once bound to the DNA end, it will start unwinding the strands and cleaving the strand ending 3' through RecB (Figure 1.16) [254, 64, 175]. It will keep on degrading DNA until it encounters a Chi site. Chi stes are specific sequences, 5' GCTGGTGG 3' in *E. coli*. RecBCD pauses at Chi sites, which will allow access

of the 5'end to the nuclease. Then, RecBCD keeps translocating but now cleaving preferentially the strand ending 5' leaving a 3' overhang (Figure 1.15). RecBCD then loads RecA before dissociating from the DNA (Figure 1.16) [28, 175].

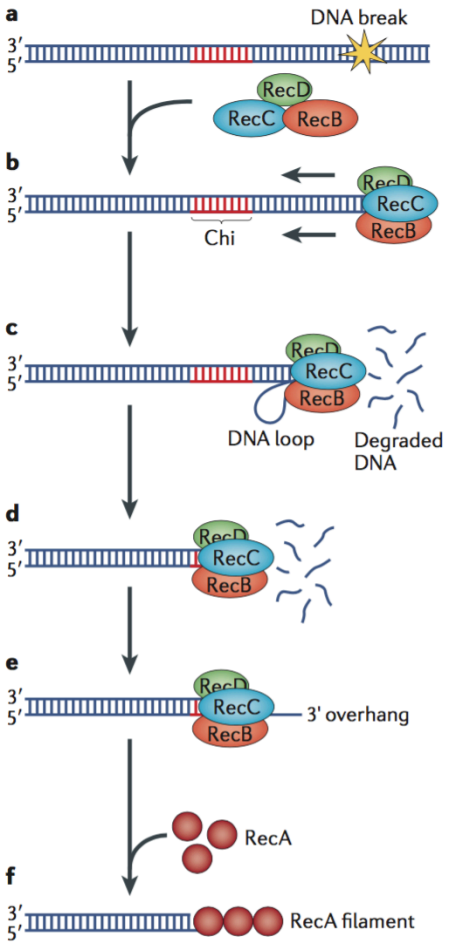


Figure 1.16: Simplified overall mechanisms for the processing of DNA ends by RecBCD.
a. DNA damage can result in a double-strand break. **b.** The RecBCD complex binds to the broken end, followed by DNA unwinding and translocation of the complex along the DNA duplex. A Chi site is denoted in red. **c.** During translocation, the RecB motor walks along a single strand in a 3'- 5' direction, whereas the RecD motor moves along the other strand in a 5'- 3' manner, giving net translocation in the same direction. As the RecBCD complex translocates, the nuclease domain of the RecB subunit degrades both DNA strands. The RecD motor translocates more quickly than the RecB motor, giving rise to a single-stranded DNA loop ahead of the enzyme. **d.** A Chi site is encountered in the 3 terminated strand. **e** The encounter with Chi induces changes that enhance degradation of the 5' terminated strand to resect the end, leaving a 3' overhang. **f.** RecBCD loads RecA onto this 3' tail and then dissociates. Figures and legends from [299].

The RecFOR pathway repairs ssDNA gaps (Figure 1.17.a). The ssDNA region is bound by SSB proteins and the role of proteins of the RecFOR pathway is to allow RecA to bypass the SSB barrier [254, 64, 175]. The first step, gap enlargement, is performed by the RecJ nuclease (Figure 1.17). It degrades ssDNA in the 5'-3' direction. It is stimulated by SSB. RecR is a ring shape protein that can accommodate dsDNA and was proposed to act as a clamp on DNA for the other proteins [175]. Two complexes RecOR and RecFR can bind DNA and recruit RecA. [231]. Even if its main role is single strand gap repair, RecFOR can act as a back up of RecBCD when RecBCD is impaired (Figure 1.17.b). In addition to RecF, RecO, RecR, the pathway to repair DSB requires RecN, RecQ and RecJ (Figure 1.17.b) [175, 15]. RecQ is a helicase that unwinds duplex DNA by translocating in the 3'-5' direction. It works together with RecJ (an exonuclease) to provide a substrate for RecFOR which then loads RecA. In Firmicutes, as some of the inactivation factors of the RecFOR are not present, it was suggested, that the RecFOR pathways might be used for DSB repair in other conditions and might be more frequent [15].

In many bacteria, RecBCD is absent but other DSB-resecting complexes repair DSB: AddAB and AdnAB [251, 299]. Several significant differences exist between AddAB and RecBCD. AddAB lacks a RecD homolog. It has one helicase and two nuclease domains compared to RecBCD which has two helicase and one nuclease domain. Each nuclease cleaves a different DNA strand [299]. AddAB recognizes a five bases Chi site in *B. subtilis*, it is also unable to unwind DNA duplexes in absence of SSB [299]. AdnAB was identified in mycobacterial species and harbours two motor-like domains and two nuclease modules [251]. Analogous to AddAB, each nuclease cleaves a different strand. AdnB helicase subunit is necessary and sufficient for duplex unwinding and contrary to AddAB, the rate of DNA unwinding by AdnAB is relatively independent of the presence of SSB [251, 299].

Synaptic and post synaptic steps

The central synaptic step of homologous recombination is achieved by RecA which catalyzes strand exchange (Figure 1.17). All of the presynaptic steps described above lead to RecA loading on ssDNA. The active form of RecA is a filament of proteins bound to ssDNA. Once formed, the RecA filament catalyzes a search for homology, then the strand invasion. This leads to the formation of heteroduplex regions where branch migration, the process by which the heteroduplex is extended, can occur [85, 54]. If the RecA filament does not encounter a homologous sequence, it lengthens, persists and induces SOS response. A second major role for RecA is regulation. RecA interacts with diverse proteins which allows control of recombination rates [85, 54].

The post-synaptic phase of homologous recombination involves either RuvABC or RecG proteins, which catalyze branch-migration and in the case of RuvABC, the cleavage of Holliday junctions (Figure 1.17). A Holliday junction is the product of a strand exchange between two homologous DNA molecules. In *E. coli*, RuvAB catalyzes branch migration [175, 15]. The RuvA tetramer plays a structural role while RuvB is an ATPase that promotes branch migration. The Holliday junction is resolved by RuvC, an endonuclease that nicks exchanged strands which are then rejoined by DNA ligase. In Firmicutes, RecU can also resolve Holliday junctions [175, 15].

The role of RecG is still not well understood partly because it has diverse functions ranging from migrating Holliday junctions to unwinding DNA/RNA hybrids via opposing RecA-mediated strand exchange. RecG is a helicase with a high specific activity of branch migration [175]. It is produced in small quantities but a 1,000- fold-lower molar concentration of RecG compared to RuvABC is required for a similar unwinding of a synthetic Holliday junction [153]. In Firmicutes, a proposed model of RecG resolution of Holliday junction involves MutS [15] : RecG catalyzes branch migration that is then resolved by MutS and concomitant ligation. RecG also has a regulatory role. It antagonizes RecA-mediated strand exchange in vitro [297].

Homologous recombination is a highly regulated process. In *E. coli*, *recX* gene is expressed downstream of *recA* in the same operon and is a RecA inhibitor. It binds to the RecA filament and promotes its disassembly. RecX is counteracted by RecFOR. Helicase UvrD limits recombination by removing RecA from ssDNA. The role of those anti recombinants might be to limit the recombination rate but also to prevent formation of a deleterious RecA filament. SbcB and SbcCD are other well known anti-recombinants as their inactivation is necessary for the RecF pathway to take place in *E. coli*. MutS2 in *H. pylori* blocks strand exchange, while MutS2 from *Thermus Thermobacter* might suppress HR through the resolution of early

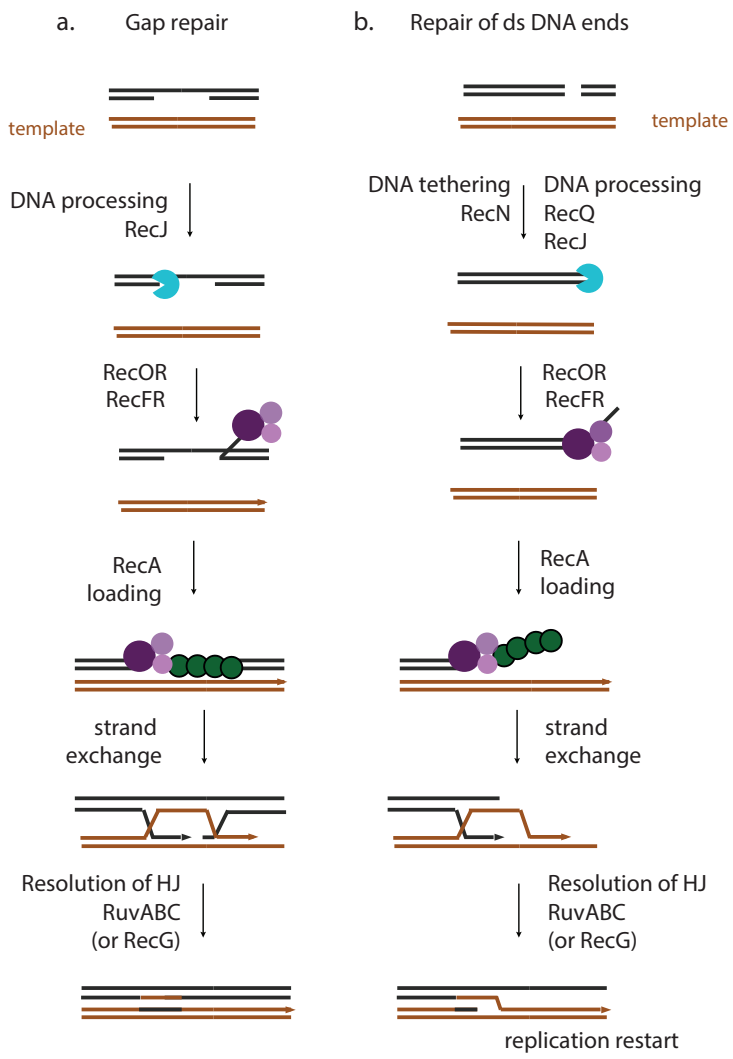


Figure 1.17: The RecFOR pathway.
a. Recombination at gaps. A gap is enlarged by the 5'-to-3' exonuclease RecJ (blue). The ssDNA region is bound by SSB (not shown). Two complexes RecOR and RecFR can bind DNA and recruit RecA. RecFR complex binds at the intersection betwen the ssDNA of the 3' and the dsDNA. Then, RecO removes SSB from the ssDNA, to allow RecA loading at the ssDNA-dsDNA junction. The RecOR pathway requires an interaction between RecO and the C-terminus of SSB but allows RecA loading without ssDNA-dsDNA junction. RecA promotes strand exchange. The Holliday junctions are resolved by RuvABC (or by a RecG-dependent branch migration mechanism in a ruvAB mutant). Black and brown lines represent the DNA strands of two homologous molecules, arrows are 3 ends. **b.** In *recBC sbcB sbcC* mutants, RecQ helicase and RecJ (blue) exonuclease provide a 3 ssDNA end onto which RecFOR (purple) loads RecA (green). RecN is required for the formation of recombinants, presumably to facilitate intermolecular interactions. Resolution requires RuvABC, and completion of the recombination reaction requires replication restart. Adapted from [175].

recombination intermediates [83]. Other proteins involved in the SOS response and DNA replication also interact with RecA [55, 15].

1.4.3 Molecular mechanisms of Non Homologous End Joining

As seen above, even if HR is ubiquitous in bacteria, it does not represent the only mean by which bacteria repair DSB. NHEJ pathway which can repair DSB without any template is present in roughly 25% of bacteria [227].

NHEJ is a two-component DNA repair pathway carried out by Ku and LigD. It is a direct ligation of processed ends (Figure 1.18). NHEJ has the advantage that it can take place at any time during the cell cycle but the disadvantage that the repair can be of low fidelity for blunt-end and complementary 5'-overhang DSBs [93]. With only two proteins it possesses all the break-recognition, end-processing, and ligation activities necessary to repair DSBs [38, 213]. Even if it was shown that only two genes are sufficient to encode a NHEJ pathway, in several examples, genomes can encode several copies of Ku [227] and different ligases can serve as back up for LigD [246].

Ku functions as a homodimer. It possesses a central core dimerization and DNA binding region that binds preferentially to double strand ends. It is also able to passively translocate along DNA [292]. LigD is an ATP-dependent ligase composed of three domains : ligase (LigDom), polymerase (PolDom) and nuclease (NucDom) [12, 291]. In some species, only the LigDom is found [213] while in others the PolDom and NucDom play important role in recognition of NHEJ DNA intermediates [60, 214].

The NHEJ pathway starts with the binding of Ku as homodimer on DNA ends (Figure 1.18). Ku then recruits LigD via the PolDom at the termini of the DSBs [292, 214]. When present, microhomologies are used to bridge the DSB, increasing the efficiency of the process and limiting the loss of genetic information. Once a stable synaptic complex is formed, the 3' ends are resected and then extended. Finally, ligD ligates the nicks [213]. Ku does not interact directly with the LigDom. Therefore it might stimulate the ligation activity by providing both free ends in the neighborhood and therefore promoting their association [213]. In certain cases, the system is complemented by additional ligases [10].

NHEJ is the only mechanism allowing DSB repair when only one copy of the genome is available. While HR will repair most breaks during exponential phase, NHEJ is expressed in *B. subtilis* in stationary phase [215]. It protects bacteria or spores that will face desiccation and dry heat [180]. Another role for NHEJ has been put forward : its involvement in stationary phase mutagenesis [203]. As NHEJ

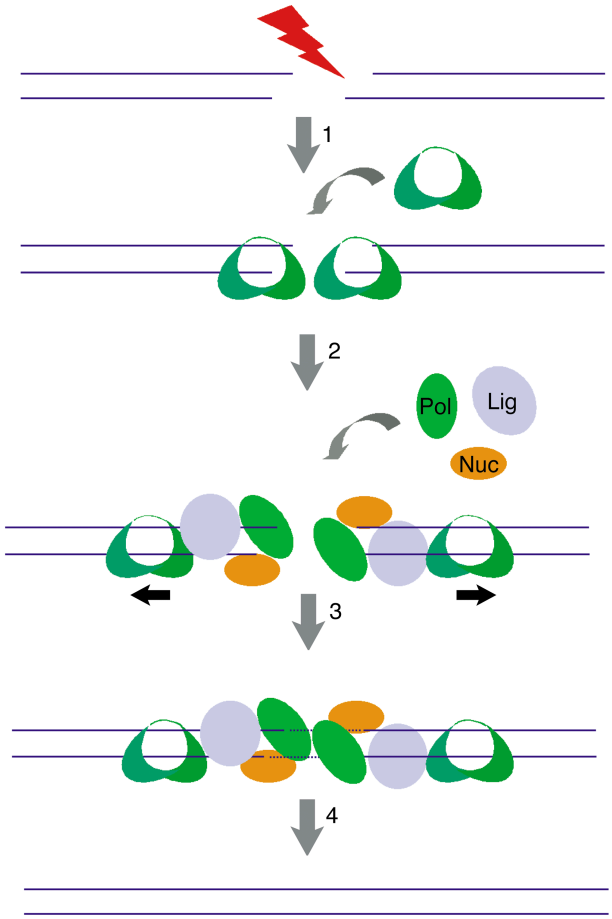


Figure 1.18: DSB Repair by the NHEJ Complex.
The Ku complex locates to the break site (step 1), where it may serve as an end-bridging and alignment factor. Following binding to the broken DNA ends, additional processing enzymes are recruited by Ku to the break site (step 2). Ku may translocate away from the ends, allowing access by other factors to the break termini. When DNA ends are non- complementary and/or are damaged, the DNA end-processing, gap- filling, and nucleolytic activities generate DNA termini, capable of being ligated, prior to ligation (step 3). Subsequently, the broken ends are joined by an NHEJ-specific DNA ligase (step 4) and the NHEJ complex dissociates. Figure and legend from [38].

repair does not rely on a template, it lacks precision when no microhomologies are present. In certain bacteria the error rates can reach 50%, thus generating diversity in stationary phase [93, 203].

1.5 Objectives : the study CRISPR-Cas systems interactions with DNA repair pathways

1.5.1 Interactions at the heart of genome genome editing technologies

CRISPR-Cas systems are mostly known because of the genome editing technologies that emerged from their use. The core idea behind those techniques is the generation of a double strand break at a specific point. As effectors complexes from Class 2 CRISPR-Cas systems only consist of a single protein, they are the most widely used. Once the double strand break is produced, it will be repaired by the host DNA repair pathways.

In mammalian cells, a vast majority of the breaks are repaired by the Non Homologous End Joining pathway (NHEJ). This will generate small indels which makes the generation of knockout mutants very easy to achieve. It is also possible to provide a template DNA at the same time as the effector and rely on homologous recombination to repair the break. This allows the introduction of a specific sequence or a desirable mutation. How the break will be repaired is therefore a central question (Figure 1.19).

In bacteria, the picture is quite different. As explained above, when a bacterial chromosome is cut, for example in autoimmunity context, the bacteria will in most cases die. Contrary to mammalian cells, it seems that the bacterial DNA repair machinery cannot keep up with CRISPR-Cas efficiency (Figure 1.18). This observation has led to the development of sequence specific anti-microbials [30, 52]. However, it has constituted a hurdle to achieve efficient CRISPR based genome editing techniques in bacteria. The first strategies developed in *S. pneumoniae*, *E. coli* and *Lactobacillus reuteri* used Cas9 mainly as a selection tool, that is to say to kill the bacteria that did not possess the mutated version of the gene [127, 195].

Efforts have been made to better understand this difference between mammalian and bacterial genome editing techniques success [57]. When introduced into *E. coli*, Cas9 cuts all copies of the chromosome simultaneously, preventing repair [57]. However, cuts in some loci in the chromosome could be tolerated. The breaks generated by Cas9 were constantly repaired by homologous recombination. The introduction of an exogenous NHEJ did not rescue a majority of the cells but led to some rare mutants exhibiting small deletions due to unfaithful NHEJ repair. This demonstrates that NHEJ can also repair Cas9-mediated breaks in bacteria, but is inefficient in this context [57].

Further studies have developed new strategies to edit the genomes of a wide range of bacterial species [185]. However, they all rely either on highly recombinant

bacteria and their endogenous DNA repair machinery or on the introduction of exogenous DNA repair pathways, and in particular phage recombinases. A better understanding of the interactions between CRISPR-Cas systems and DNA repair pathways while encompassing the complexity and diversity of both defense and repair systems seems essential to improve existing microbial biotechnologies.

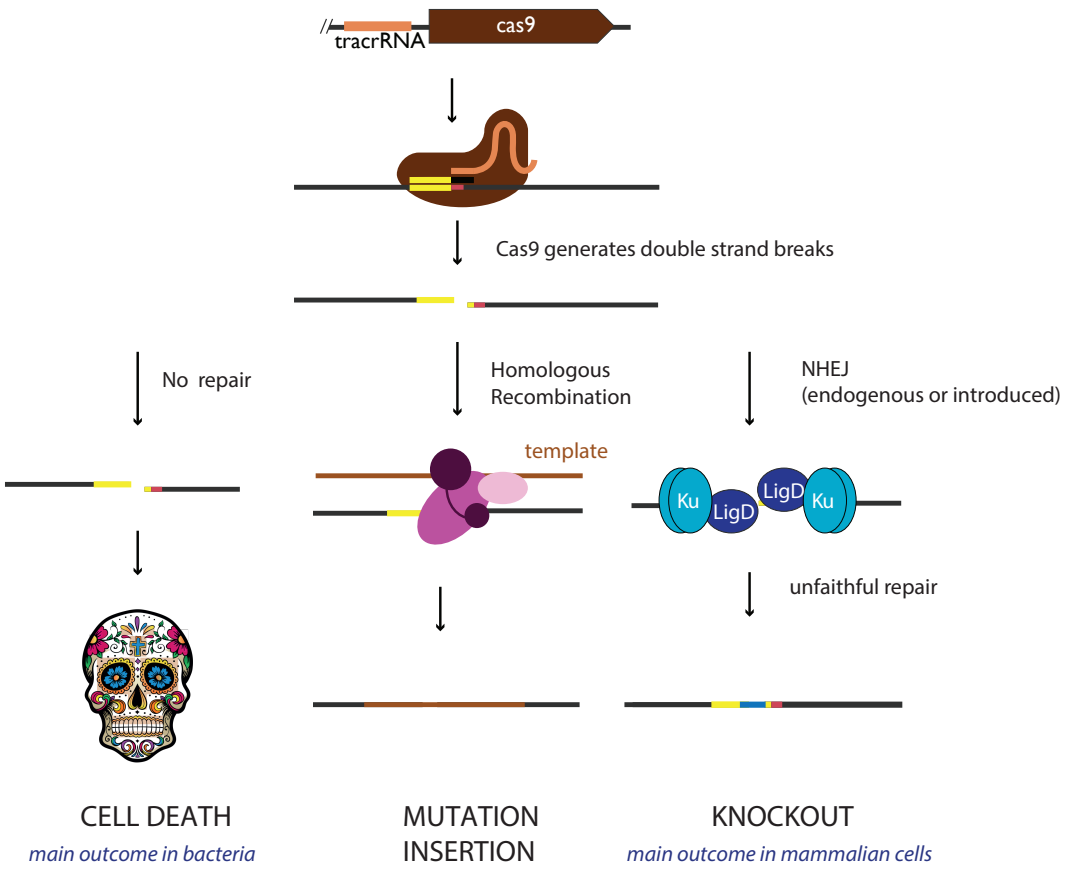


Figure 1.19: Interactions between CRISPR-Cas systems and DNA repair pathways at the heart of genome editing techniques
Cas9 is used to generate a double strand break which will then be repaired by the host DNA repair machinery. Following the DSB, three outcomes are possible : cell death if no repair is achieved, mutation or insertion of a new sequence through homologus recombination with a given template, gene inactivation through unfaithful repair by NHEJ. In mammalian cells, NHEJ repair is the most frequent outcome while cell death prevails in bacteria.

1.5.2 Interactions between CRISPR-Cas systems and DNA repair proteins in bacteria

The interplay between CRISPR-Cas systems and DNA repair proteins is important not just in genome editing experiments but also at different stages of CRISPR-Cas adaptation and immunity [137]. DNA repair proteins have proven essential for specific adaptation steps in some bacteria. As detailed in the section dedicated to the molecular mechanisms of adaptation, RecBCD provides prespacers for type I-E adaptation machinery [149] demonstrating a clear synergistic interaction between the two systems. Following this finding, other experiments have underlined the role of other DNA repair proteins in adaptation [122, 110]. RecG and PriA contribute to primed adaptation in type I-E and I-F CRISPR-Cas systems [122, 110].

Recently, further evidence of a link between CRISPR-Cas immunity and DNA repair was introduced by the demonstration of a translational coupling of the two processes in *Sulfolobus islandicus* [152]. More specifically, a known activator of type I-A CRISPR-Cas system in *Sulfolobus islandicus*, Csa3a was found to activate the expression of adaptation genes, CRISPR RNAs and DNA repair genes. It resulted in enhanced crRNA biogenesis and spacer acquisition. The authors propose that the archaeal DNA repair proteins involved play a similar role as RecBCD [152].

Other studies have hypothesized a role for Cas proteins in DNA repair. First, CRISPR arrays were initially thought to be involved in replicon partitioning [183] and Cas proteins to be a DNA repair system for thermophilic archaea and bacteria [159]. The main argument for *cas* genes being part of a repair system was their predicted functions including DNA helicase, exonuclease, DNA polymerase, and the similarity of one of the protein to RecB [159]. Further studies brought up experimental evidence of the role of cas proteins in DNA repair [300, 16]. Expression of *cas* genes in *Pyrococcus furiosus* increases following exposure to gamma-radiation [300]. A study also found that the deletion of Cas1 or of the CRISPR array in *E. coli* led to an increased sensitivity to DNA damage and impaired chromosomal segregation [16]. Moreover they show that Cas1 interacts physically and genetically with essential DNA repair proteins: RecB, RecC and RuvB [16].

Finally, it is interesting to note the functional and structural similarities between Cas proteins of unknown function and DNA repair proteins. The role of accessory proteins in adaptation for specific subtypes of CRISPR-Cas system has not been elucidated. Typically, type II CRISPR-Cas systems are composed of three core genes *cas9*, *cas1* and *cas2*. Type II-A and II-B harbor additional genes respectively *csn2* and *cas4* while type II-C only only carries the three core genes [161]. Both Csn2 and Cas4 have been linked to the adaptation process. Csn2 interacts with Cas1 and Cas2 and adaptation was abolished in *S. thermophilus* when *csn2* was deleted [19, 134]. Cas4 is necessary for type I-B priming in *H. hispanica* [150]. It

interacts with Cas1-Cas2 fusion protein in *Thermoproteus tenax* type I-A [216] and fusion of Cas4 and Cas1 are found in several systems suggesting a general role in adaptation. Moreover, a Cas4-like protein found in *Campylobacter* bacteriophages can activate spacer acquisition in *Campylobacter jejuni* which harbours a type II-C system [113]. Both Cas4 and Csn2 share similarities with proteins involved in DNA repair. Csn2 multimers bind to linear dsDNA free ends and can then translocate along the DNA[14, 147, 138, 74]. The DNA binding properties of Csn2 are similar to the binding properties of Ku, a protein involved in NHEJ pathway. Cas4, like Csn2 forms a ring-shaped structure. It harbours a RecB domain similar to AddB [313, 148]. The role of these accessory proteins as well as the importance of the host machinery during adaptation underline the complexity and diversity of the mechanisms involved in spacer acquisition. Therefore, the study of the interactions between CRISPR-Cas systems and DNA repair pathways has brought essential knowledge about CRISPR biology and further work on these interactions might bring new critical understanding to how CRISPR-Cas systems function and evolve.

My thesis aims at understanding how CRISPR-Cas systems interactions with DNA repair pathways also shape CRISPR-Cas distribution in bacterial genomes. First, I examine precisely the distribution of CRISPR-Cas systems (Chapter 3), then I present an analysis of co-occurrence patterns of CRISPR-Cas systems and DNA repair pathways (Chapter 4) and finally I introduce experimental evidence of a negative interaction between a CRISPR-Cas system and a DNA repair pathway (Chapter 5). My findings give insights on the complex interactions between CRISPR-Cas systems and DNA repair mechanisms in bacteria and contribute to explain the scattered distribution of CRISPR-Cas systems in bacterial genomes.

Box 1: Major points of the Introduction

- CRISPR-Cas systems are an **adaptive immune system** of bacteria and archaea. CRISPR-Cas systems are organized around two components : **a cluster of *cas* genes** and a **CRISPR array** composed of spacers and repeats. They works in two steps : **adaptation** (acquisition of new spacers) and **immunity** (targeted degradation of foreign genetic element).
- CRISPR-Cas systems are **extremely diverse**. Classified in two classes, six types and 21 subtypes, this variety reflect a **diversity of molecular mechanisms** by which CRISPR-Cas systems achieve adaptive immunity.
- CRISPR-Cas systems are **widespread** in bacteria thanks to **horizontal gene transfer**. However, they **remain quite rare** for an immune system and present a **scattered distribution**.
- Several hypothesis have been put forward to explain this relative scarcity: **ecological factors** including phage diversity and abundance which tend to favor other immune systems over CRISPR-Cas systems, **autoimmunity** , **limiting horizontal gene transfer** or the **cost of maintenance** of such systems.
- The goal of this PhD was to study a new hypothesis to explain the relative scarcity of CRISPR-Cas systems: **the genetic background would play an important role in the process of maintaining a CRISPR-Cas system after its transfer**. To test this hypothesis, we decided **to focus on one essential bacterial function: DNA repair**; mainly because DNA repair and CRISPR-Cas systems share the same substrate and DNA repair pathways could repair breaks generated by CRISPR-Cas systems and thus limit CRISPR-Cas efficiency.
- In bacteria, D**NA breaks are mended mainly by two pathways**: **Homologous Recombination** (HR) and **Non Homologous End Joining** (NHEJ). Homologous recombination requires an intact template while NHEJ does not require a template but is error prone.
- **Interactions between DNA repair pathways and CRISPR-Cas systems** are at the **heart of CRISPR based technologies** where breaks generated by Class 2 effectors are repaired by endogenous DNA repair pathways. In bacteria, **several synergistic interactions have been reported** such as the dependency of type I-E systems on RecBCD for acquisition of new spacers.

Chapter 2

Methods

This chapter introduces methods for the study of CRISPR-Cas systems. Approaches to detect CRISPR-Cas systems in bacterial genomes and study their interactions are presented, followed by an overview of the experimental procedures to study the activity of these systems.

2.1 Detecting CRISPR-Cas systems in bacterial genomes and their interactions

The detection of CRISPR-Cas systems is usually made separately for its two main components, the Cas clusters and the CRISPR array, using different approaches. So far, no precise method has been developed to associate a CRISPR array with a cognate Cas cluster. Chapter 3 of this thesis is dedicated to this question.

2.1.1 Detecting Cas clusters : Macsyfinder and the first version of CAS-Finder

Tools to detect Cas clusters are based on protein profiles. Protein profiles are probabilistic models built from the information contained in protein alignments. They allow more sensitive identification of distant homologs than classical pairwise sequence-search approaches [70]. 396 protein profiles for *cas* genes were built when the classification of CRISPR-Cas systems was carried out[161]. The catalog of *cas* genes was then expanded by using metagenomics data, generating 130 new profiles [315]. The method I used to detect Cas clusters makes use of these protein profiles.

Automated annotation of Cas clusters is difficult because of the diversity and the organization of Cas and CRISPR loci. Makarova et al proceeded in two steps when they defined CRISPR-Cas classification: they first detected the Cas clusters and then typed them. The detection was itself a two steps process to limit the rate of false positives while trying to encompass the diversity of CRISPR-Cas systems

at the same time [161]. Anchor *cas* genes were first detected by using 369 custom protein profiles and a high similarity threshold. The search was then extended to the neighborhood for other potential *cas* genes with a less conservative threshold. Following the detection, the assignation of a specific subtype to a CRISPR-Cas locus was achieved in two steps. First, signature proteins were used (such as Cas3 for type I, Cas10 for type III) followed by the analysis of the other genes present in the locus. If more than 2/3 of the genes were associated to one subtype, it was classified as such [161].

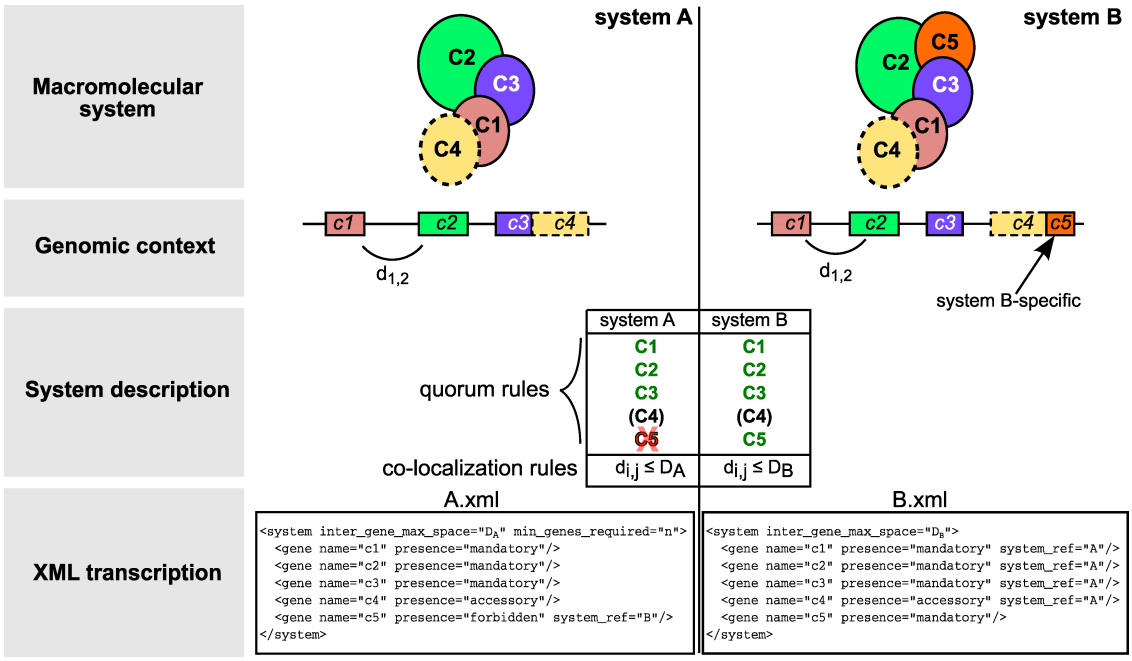


Figure 2.1: MacsyFinder models.
A macromolecular system is typically encoded in one or a few different loci. To model and differentiate two similar systems (here A and B), the user can specify homologous components (C1-C4), mandatory ones (C1-C3), forbidden ones (C5) or accessory ones (C4). These rules are transcribed in a computer-readable model in xml format. Components correspond to proteins. Figure from [1].

However, this detection and further typing was performed on a specific dataset and is not easy to reproduce on any given genome or dataset. Recently, a webserver that uses Cas protein profiles to search for Cas systems offers to type them using signature proteins [46]. However, it does not take into account the architecture of the locus and does not allow subtype annotation.

The Rocha lab followed another approach to create CasFinder, a software to detect and type CRISPR-Cas systems [1]. CasFinder is based on MacsyFinder, a program to mine genomes for macro molecular systems. MacsyFinder provides a

flexible framework to model the properties of molecular systems including information on the genetic architecture and the minimally sufficient number of components. The components defined in the models are searched by sequence similarity using HMM protein profiles. The assignment of hits to a given system is decided based on their compliance with the content and organization defined in the model (Figure 2.1) [1].

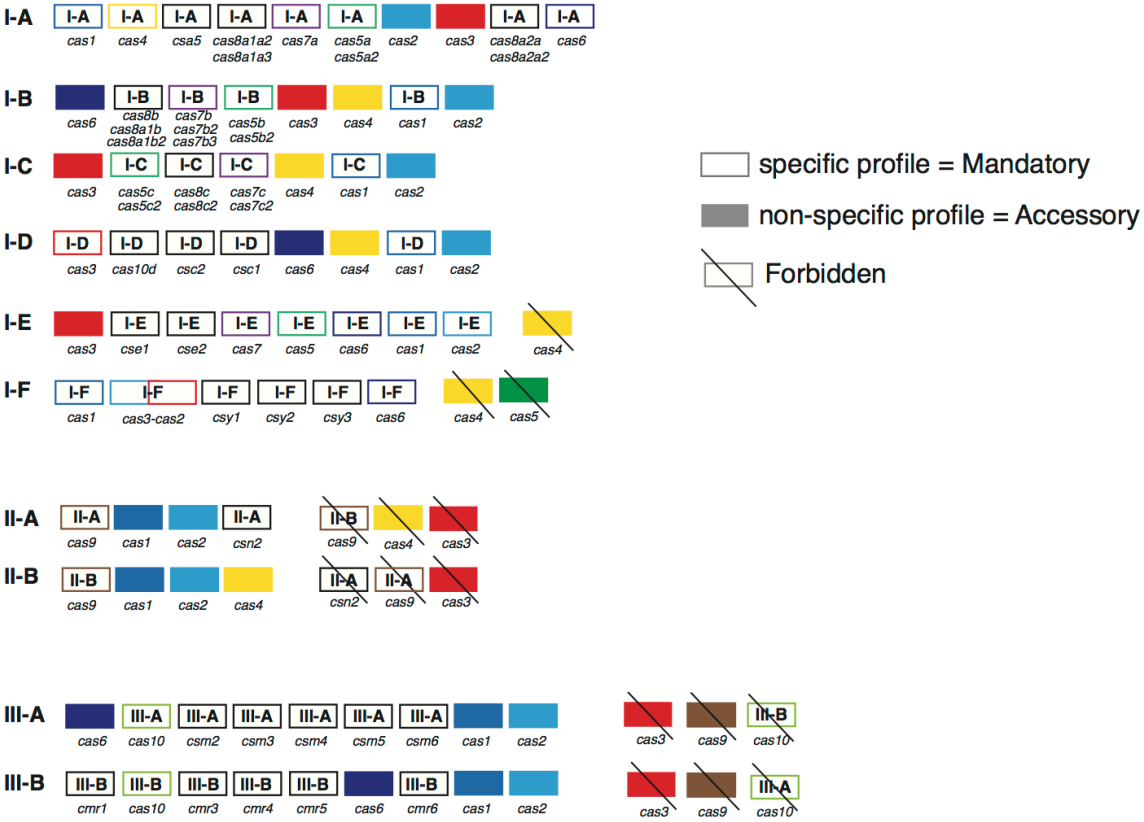


Figure 2.2: Models of the first version of CasFinder.
Rectangles with colored outlines correspond to signature proteins used to subtype the different systems. They are defined as mandatory in the models. Fully colored rectangles are proteins present in several systems which cannot be used to subtype. They are therefore defined as accessory in the models. Crossed rectangles correspond to forbidden proteins. They are used to identify systems subtypes, more specifically to distinguish between several close subtypes (example csn2 and cas4 in type II-A and II-B). Figure from [1].

The first version of CasFinder had a set of models to identify and class CRISPR-Cas systems in three types (I,II,III) and ten subtypes (I-A, I-B, I-C, I-D, I-E, I-F, II-A, II-B, III-A, III-B). These models are represented on Figure 2.2. Briefly, they are based on the presence or the absence of signature proteins for each subtype. Proteins that are present in different systems are defined as accessory to allow

their detection in the locus without requiring them for subtype identification. Forbidden proteins (ie proteins that cannot belong to a defined subtype) are also used to subtype. Three levels of precision (subtype, type and detecting generally Cas clusters) were developed to deal with the tradeoff between the detection of a maximum of Cas clusters and a precise identification of each cluster. The general Cas model was designed to detect any cluster of three Cas proteins or more.

2.1.2 Updating Cas-Finder

With the publication of the new classification of CRISPR-Cas systems by Makarova and colleagues in 2015 [161], CasFinder needed an update to detect the three new types and the four new subtypes of CRISPR-Cas systems. Moreover, as 394 new profiles had been published with the classification, the models and the protein profiles had to be revised. This was part of the methodological work I developed in the course of my PhD.

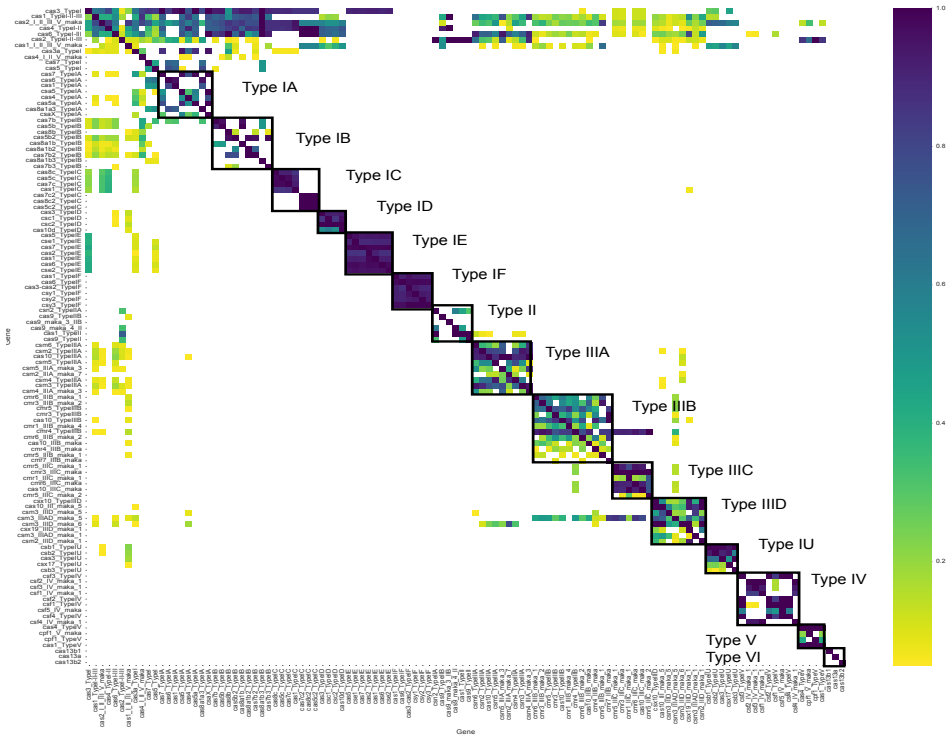


Figure 2.3: Frequency of co-occurrence between Cas proteins present in clusters detected with the subtyping models of the new version of CasFinder. Matrix was normalized by the maximum of each column. The higher the frequency is, the darker the colour is. Only frequencies above 5% were represented, others are in white.

The first step was to develop new models for the new types and subtypes. These were designed to match the described classification. In the existing models, the new protein profiles were incorporated. This constituted a first basis for the work.

However, many issues arose at that point. First, the large number of different profiles caused the program to be too slow. It was therefore necessary to reduce profile diversity. To do so, I ran CasFinder with all the profiles to identify the ones used most frequently. Profiles that rarely matched proteins in the database (fully sequenced genomes available in November 2013) were removed from the definition and only the profiles with the most hits were kept for the detection of one protein (e.g. Cas1). The final choice of profiles was made so that all the loci could still be detected.

Second, it became apparent that some profiles were not strictly associated with one subtype (for example Csm2 from type III-A matched other type III systems), and therefore could not be confidently used to type systems. Non-specific profiles were detected by computing the co-occurrences between profiles, and grouping them into clusters (Figure 2.3). Profiles that could not be confidently assigned to any cluster were either removed or defined as accessory in the models. As a result, the final model relies on few signature proteins for the identification of each subtypes. Finally, I had to minimize inconsistencies between the different levels of detection (Cas clusters, type or subtype). By examining the disagreements, I made adjustments to the definitions to limit the number of missed clusters while making more precise annotations.

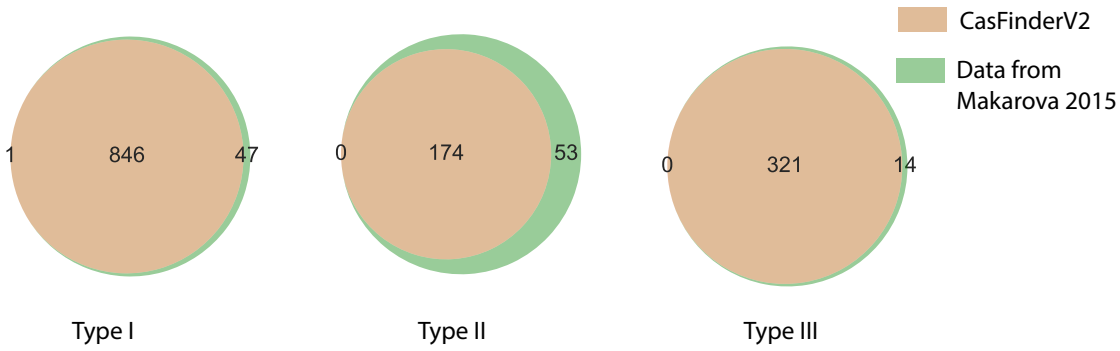


Figure 2.4: Comparison of CasFinderV2 with the detection published for the updated classification of CRISPR-Cas systems [161].
The figure represents Venn diagrams with te results of the detection of the three main types of CRISPR-Cas systems. Numbers represent the number of genomes where a specific type was detected.

The new version of CasFinder is available on the galaxy portal of the Institut Pasteur. Xml models are presented in Annexe 1. Relative to the previous version,

it contains 41 new protein profiles, for a total of 117 profiles. A comparison between the detection of CasFinderV2 and the one by Makarova and colleagues revealed few differences (Figure 2.4). CasFinder V2 is more conservative than the detections made by Makarova and colleagues which identified 47 Type I, 53 Type II and 14 Type III, not found by CasFinderV2. This can be attributed to the fact that Makarova detected clusters with only 2 Cas proteins while CasFinderV2 requires at least 3. The detections performed in my thesis used CasFinderV2.

2.1.3 Detecting CRISPR arrays

Several pieces of software to detect CRISPR arrays have been developed: CRISPR Recognition Tool (CRT), CRISPRFinder, PILER-CR and more recently CRISPRDigger [32, 96, 71, 89]. They all look for repeats and then identify a CRISPR array based on repeat size, identity and spacing. Several additional steps can help limit the CRISPR array and determine the exact repeat and spacers sequences.

In the analysis provided in this thesis, we used CRT [32]. CRISPRFinder was not available as a standalone version but only as a webserver, CRISPRDigger had not been released yet and authors of CRT claimed that CRT was faster than PILER-CR. A recent analysis made for the release of CRISPRDigger showed that the different pieces of software performed relatively similarly [89].

CRT is based on finding a series of short exact k-mers separated by spacers of similar size. These exact k-mer matches are then extended to the actual repeat length. k should be smaller than the length of the shortest repeat [32]. Once a cluster of repeats is identified, it is kept only if it meets specific requirements: 1) repeats must fall within a specific size range, 2) the first spacers must but different from one another, 3) spacers must be of similar size. Finally, both flanks of the arrays are checked to find more divergent repeats. This is important because the last repeat in CRISPR arrays is often degenerated [32, 96].

By default, the maximal length of repeats in CRT is 38. However, some bacteria encode CRISPR arrays with larger repeats so we set that parameter to 50. This allowed the recovery of 154 new CRISPR arrays on a total of 7122 detected in 5563 complete bacterial genomes. To raise the quality of detection, we further tuned one parameter: the size variation of the spacer (Figure 2.5). The variability of spacer size in an array can be measured by the coefficient of variation (=standard deviation/mean). For CRISPR arrays, this coefficient is expected to be low, as spacers are integrated through mechanisms that ensure constant size [124]. As expected, the number of detected arrays drops when the coefficient of variation rises. We then wanted to define a threshold above which detected arrays would be considered false (ie the detected repeats would not correspond to CRISPR arrays). To do so, we used the presence of *cas* genes in the neighborhood (here 10kb) as

a measure of probability of the encoded repeats being real CRISPR arrays. The observed frequencies drop after a coefficient of variation of 19 and are close to zero after 28. We therefore chose to remove from the dataset detected arrays with a coefficient of variation superior to 28 because we were confident that they were not CRISPR arrays. This step removed 318 arrays representing 4.4% of the dataset. The total number of arrays detected remaining was 6958.

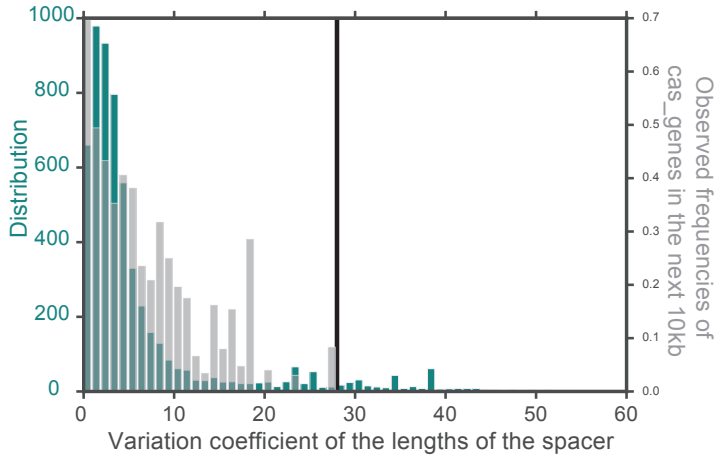


Figure 2.5: Detection of CRISPR-arrays : removing false elements.
In green, the distribution of the coefficient of variation of the size of spacers of CRISPR arrays, in grey the observed frequencies of *cas* genes within 10kb . The chosen coefficient of variation threshold of 28 is marked by a vertical line.

2.1.4 Detecting interactions in bacterial genomes

Statistics and phylogeny

I will detail here the statistical methods used to study the associations between genes or groups of genes in bacterial genomes, in particular between CRISPR-Cas systems and DNA repair pathways, the central goal of this work. The key element in this analysis is to assess whether the number of co-occurrences between two elements is different from the one expected under a null model. Once the two systems are detected, the expected value of the number of co-occurrences can be computed as the product of the marginal row and column totals divided by the grand total of the contingency tables. We can then test whether the observed number of co-occurrences is different from the expected one with a Fisher exact test.

This test assumes independence between individuals (bacterial strains in this case). However, bacterial strains are part of a hierarchically structured phylogeny and often cannot be regarded as independent. Hence, one must account for the

phylogeny to make a more accurate statistical analysis. The first method to correct statistics by the phylogeny was introduced by Felsenstein and concerned continuous traits [79]. As the interactions presented here concern discrete traits *i.e.* the presence or absence of specific systems, the contrast method cannot be applied directly. Instead, we used here a method adapted to discrete characters developed by Pagel [200]. The method uses a continuous-time Markov model to characterize evolutionary changes along each of the branches of a phylogenetic tree. It compares the goodness of fits of two different models to the observed data set : one where the two characters are treated as evolving independently and a more parameter-rich one where the characters evolve in a correlated fashion. The two models are compared using likelihood ratio test [200].

Building phylogenetic trees for Firmicutes and Proteobacteria

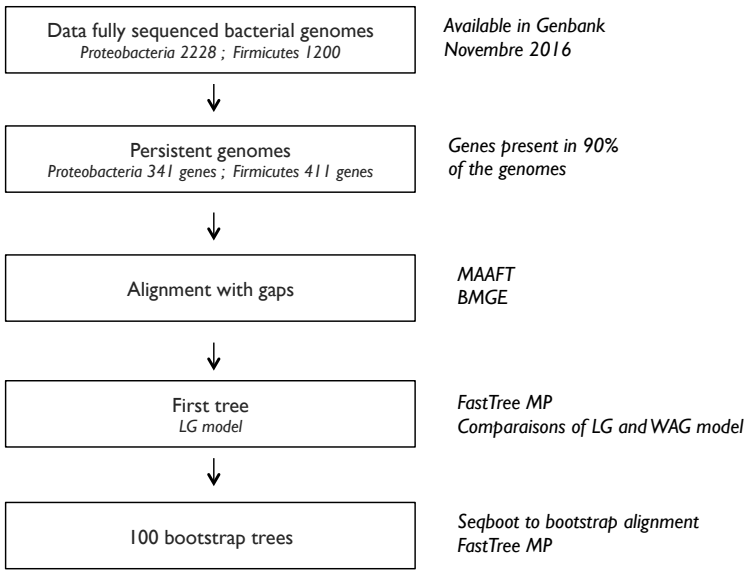


Figure 2.6: Diagram of the method used to build phylogenetic trees

The first step to perform such controls is to build phylogenetic trees. As our analysis include a very large number of species this led to specific challenges. When considering a large number species, only a small number of genes can be found that are common to all (the core genome), leading to trees of poor resolution. Therefore, we did not attempt to build a tree of all bacteria but performed most of the analysis on specific clades that had enough genomes encoding CRISPR-Cas systems (Firmicutes and Proteobacteria). This increased the robustness and accuracy of the statistical analysis. Since bacteria with a genome size inferior to 1Mb do not encode CRISPR-Cas systems, they were discarded from the analysis. Hence, our final sample has 1189 genomes of Firmicutes and 2897 of Proteobacteria.

One way of simplifying the dataset is to work at the species level and not at the strain level. However, contrary to DNA repair pathways which are usually common to all the strains in a given species, CRISPR-Cas systems are not persistent, making it hard to associate the presence or absence of a CRISPR-Cas system to a given species. It thus appeared that an analysis at the strain level was necessary.

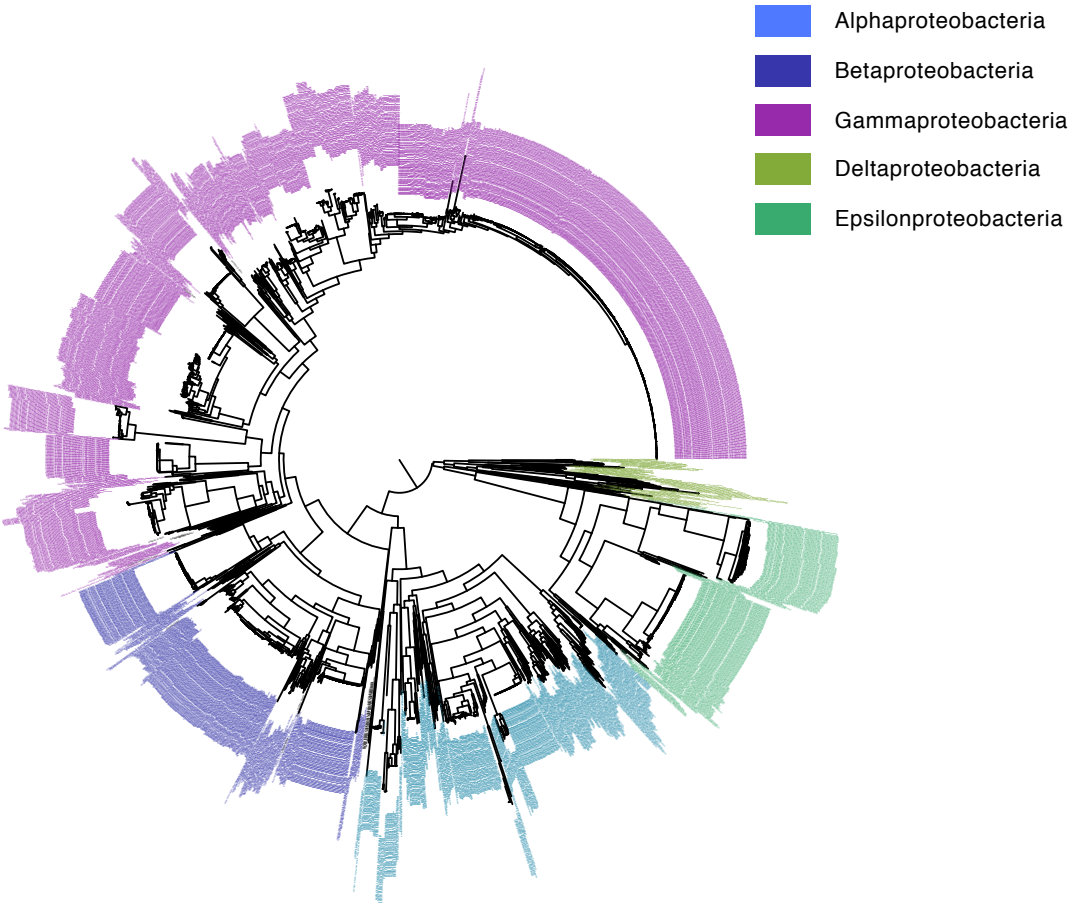


Figure 2.7: Phylogenetic tree of Proteobacteria

Reconstructing trees with many diverse genomes is computationally challenging. I first tried to build a 16S RNA tree, but the resolution was not sufficient. I then computed the core genome of all or Firmicutes and of Proteobacteria to build a tree based on more characters. However, given the number of genomes, the core genomes were reduced to less than 10 genes, greatly limiting the number of positions that could be aligned. We therefore chose to analyse all families of orthologous genes that were present in more than 90% of the genomes (Figure 2.6). To identify this persistent genome, a list of orthologs was identified as reciprocal best hits using end-gap free global alignment, between the proteome of a pivot and each of the other strain's proteomes. Escherichia coli K12 MG1655 and Bacillus

subtilis str.168 were used as pivot for Firmicutes and Proteobacteria. Hits with less than 37% similarity in amino acid sequence and more than 20% difference in protein length were discarded. The persistent genomes of Firmicutes contained 411 families and that of Proteobacteria 341 families.

For each gene of the persistent genome, proteins were aligned using MAFFT (v.7.205 with default options) and BMGE (v1.12 with default options). We then concatenated the multiple alignments and replaced missing genes with stretches of "_". The addition of "_" does not impact phylogeny reconstruction [80]. The very large size of the multiple alignment, which for Proteobacteria consisted of 341 genes and 2897 strains, precluded the use of most of the available programs to reconstruct trees (due to the requirements in terms of memory and computational time). The trees were therefore computed with FastTree, a program specifically made to analyse such large datasets [218]. The likelihoods of the tree under two models LG and WAG were compared. In both cases, the LG model was the one that minimized the Aikake Information Criterion. Hence, the trees were inferred using this model and we made 100 bootstraps using phylip's SEQBOOT to generate resampled alignments and the -n -intree1 options of FastTree.

Controlling with the phylogeny

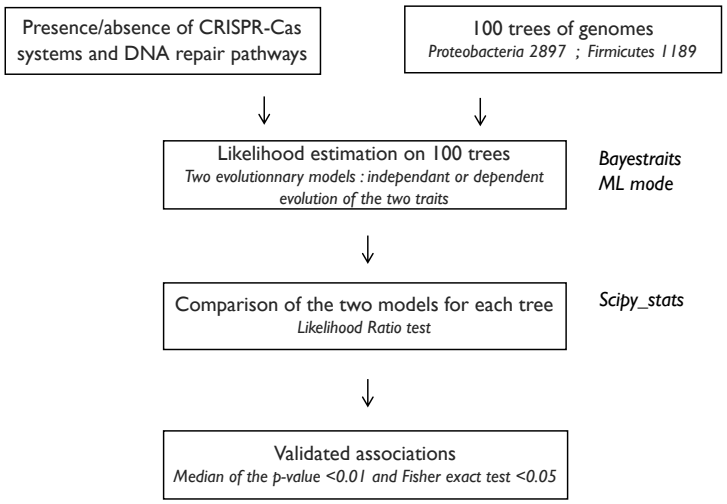


Figure 2.8: Diagram of the method used to detect significant associations between two systems in bacterial genomes

To perform the control by the phylogeny; we employed a program called BayesTraits (V3) designed to analyse the correlated evolution of discrete characters in a phylogeny [201]. The program can estimate the likelihood of the presence/absence pattern of discrete traits given a specific phylogeny. The likelihood is estimated

under two models: one where it is hypothesized that the discrete traits evolved independently (which defines four transition rates for binary data) and one where the characters have evolved in a correlated manner (which defines eight transition rates, going from 1 to zero for both characters given that the other character is equal to zero or one). The likelihood of the data given the model and the phylogenetic tree is the product of the likelihoods for each character which is based on the product over all the branches of the tree of the appropriate probability for each branch derived from the Markov processes. The likelihood of these two models for a specific dataset can be compared using a likelihood ratio test. The likelihood ratio test takes into account the number of parameters (4 and 8) and tests the null hypothesis that the most complex model is not better than the simpler one given the difference in the number of parameters. If the null hypothesis is rejected, the more complex model - correlated evolution - can be regarded as significantly better.

The phylogeny of large clades like Firmicutes and Proteobacteria leaves certain parts of the tree poorly resolved. To take into account uncertainty in the phylogenetic reconstruction, we computed separate likelihood-ratio tests on the 100 trees provided by the bootstrap analyses. We considered that an association was significant if the median of those 100 likelihood ratio tests was inferior to 0.01 and if the p-value of the Fisher Exact test was inferior to 0.05.

2.2 An introduction to experimental approaches to study CRISPR-Cas systems

The introduction to experimental approaches to study CRISPR-Cas systems is presented as a chapter of a book of the Methods in Molecular Biolgoy series : Methods in Horizontal Gene Transfer. It gives an overview of how to examine CRISPR-Cas activity by taking the example of the study of the type II-A CRISPR-Cas system from S. pyogenes in S. aureus.

Methods for the analysis & characterization of defense mechanisms against HGT II: CRISPR

Alicia Calvo-Villamañán[1], Aude Bernheim[1,2,3], David Bikard[1]
[1] Synthetic Biology Group, Microbiology Department, Institut Pasteur, Paris 75015, France
[2] Microbial Evolutionary Genomics, Department of Genomes and Genetics, Institut Pasteur, Paris 75015, France
[3] AgroParisTech, Paris 75005, France

i. Summary/Abstract

CRISPR-Cas systems provide RNA-guided adaptive immunity to the majority of archaea and many bacteria. They are able to capture pieces of invading genetic elements in the form of novel spacers in an array of repeats. This can then be used as a memory to destroy incoming DNA through the action of RNA-guided nucleases. This chapter describes general procedures to determine the ability of CRISPR-Cas systems to capture novel sequences and to use them to block phages and horizontal gene transfer. All protocols are performed in *Staphylococcus aureus* using type II-A CRISPR-Cas systems. Nonetheless, the protocols provided can be adapted to work with other bacteria and other types of CRISPR-Cas systems.

ii. Key words

Adaptation, Adaptive Immunity, CRISPR-Cas, Horizontal Gene Transfer (HGT), Interference, Mobile Genetic Elements (MGE), Protospacer, Protospacer adjacent motif (PAM), Spacer

1. Introduction

1.1 – CRISPR-Cas systems and their classification

The CRISPR-Cas systems (Clustered Regularly Interspaced Short Palindromic Repeats, CRISPR associated) are RNA-guided, adaptive immune systems found in many bacteria (~50%) and most archaea (~90%) [1]. The genomic structure of the system was first identified in 1987 [2], but its role as an adaptive immune system against mobile genetic elements (MGEs), and phages in particular, was only proposed in 2005 [3–5], with experimental evidence published in 2007 [6]. These systems were first hypothesised to be a force that counteracts horizontal gene transfer (HGT) in 2008, when a study of *Staphylococcus epidermidis* revealed the presence of a CRISPR-Cas system that targets the nickase gene present in the majority of the staphylococcal conjugative plasmids [7]. CRISPR-Cas systems are able to destroy target DNA elements whether carried by the chromosome, phages or plasmids [8] and can thus block horizontal transfer regardless of the mode of entry: injection by phage particles, plasmid conjugation or natural transformation [9].

A CRISPR consists in an array of short palindromic direct repeats, separated by non-repetitive sequences, called spacers, which can be easily identified through bioinformatics tools [10, 11]. The spacers provide a genetic memory of past encounters with exogenous genetic elements. The size of individual CRISPR arrays can vary from a single spacer to more than 300, and genomes can carry more than 10 CRISPR arrays. While most CRISPR arrays are found in association with *cas* genes at the same locus, *cas*-less CRISPR arrays are also frequently identified. These CRISPR can either rely on *cas* genes present at a distant locus in the genome or be inactive. Different CRISPR-Cas systems present distinctive combinations of *cas* genes frequently found in operons. In fact, it is based on the difference between the architecture and protein content of the *cas* operons that these systems are classified [1].

In the last years, a great variety of different CRISPR-Cas systems have been described in both bacteria and archaea. Obtaining a straightforward classification of these systems is hard, since they show complex dynamics. Their evolution is very fast, involving changes such as rearrangements of the *cas* operon, horizontal transfer of complete *loci,* or modular parts [1]. We provide here a brief overview of their classification but recommend reading the following review for more detailed information [1]. CRISPR-Cas systems are divided into two broad classes: Class 1 systems, which possess multi-subunit effector complexes, and Class 2 systems, in which all functions of the effector complex are carried out by a single protein [12].

Class 1 systems include three types, type I, type III and type IV. Type I CRISPR-Cas systems contain the signature gene *cas3*, which encodes a single-stranded DNA helicase that acts both on dsDNA and RNA-DNA duplexes [13–15].Type III CRISPR-Cas systems possess the signature gene *cas10*, which encodes a multidomain protein. Lastly, type IV CRISPR-Cas system present a minimalistic architecture and their signature gene, *csf1* [1].
Class 2 systems include types II, V and type VI. Type II systems' signature gene is the well-known *cas9.* This protein has been studied in detail due to its use as a biotechnological tool [16]. Type V systems include the *cpf1* gene, a functional analogue of *cas9* [17]. Some subtypes include the gene *c2c1* instead [18, 19], which is distantly related to *cpf1*. Lastly, type VI is a recently described novel type that includes systems able to target not only DNA, but also RNA. Their signature gene is *c2c2.* [12, 18, 19].

The different types of CRISPR-Cas systems show a complex distribution amongst bacteria and archaea phyla. Overall, type I systems are the most commonly found, accounting for 50% of the CRISPR-Cas systems found in both archaea and bacteria [1]. Type II systems are only found in bacteria, but correspond to 10% of their CRISPR-Cas systems. Finally, type III make up 20-25% of the CRISPR-Cas systems across both archaea and bacteria [1].

1.2 – CRISPR-Cas immunity

CRISPR-Cas immunity is divided into three phases. The adaptation phase in which short sequences from the invader, also called protospacers, are integrated as novel spacers in the CRISPR array. The expression phase, in which the CRISPR array is transcribed and processed into small crRNA. The immunity or interference phase, in which the crRNA guides a complex of Cas proteins to destroy target nucleic acids. These steps are common to all CRISPR types, but the mechanism employed and the nature of the target can be different.

During adaptation, pieces of the invader's DNA are captured by a complex of Cas proteins preferentially from broken DNA or free DNA ends [20–23]. They are then processed and transported to the CRISPR array to be integrated in a process that will create a novel spacer and repeat [24–26]. Protospacer selection is generally guided by the presence of a specific motif, called a proto-spacer adjacent motif (PAM), next to the target sequence [27, 28]. Depending on the type of CRISPR system the PAM can be found on the 3' or the 5' side of the protospacer/target, and its sequence and size can vary. The PAM sequence from many CRISPR-Cas systems have already been identified, and methods to identify PAM motifs have recently been reviewed elsewhere [29, 30]. Note that type III systems do not rely on a PAM motif. The mechanisms involved in novel spacer selection by these systems largely remains to be investigated. It is also important to note that two modes of spacer acquisition are described in

the literature, depending on whether there is already a record of the MGE in the CRISPR array or not. When there is already a pre-existing spacer matching the MGE in the array, even imperfectly, the acquisition is called primed [28, 31]. On the other hand, when there is no spacer in the array matching the MGE, the spacer acquisition is said to be naïve [32]. Primed acquisition occurs at a much higher frequency than naïve acquisition, presumably thanks to the generation of DNA breaks by Cas proteins guided towards the MGE [33].

During the expression phase, the CRISPR array is first transcribed into a pre-crRNA which may contain a series of hairpins due to the CRISPR's palindromic repeats. The pre-crRNA is then processed either by Cas enzymes or host's RNAses into smaller units that correspond to a single spacer flanked by partial repeats.[34–37] In type II systems, processing is dependent on a trans-activating CRISPR RNA (tracrRNA), which hybridizes to the repeat sequences in the pre-crRNA [35].

During the interference step, the mature crRNA, binds to a complex of Cas proteins, scans the DNA in the cell and locates the corresponding target [38, 39]. In CRISPR-Cas systems of type I, II and V, the PAM motif is first located. Then DNA is unwound and base pairing is established between the crRNA and the target DNA strand. Binding in the region next to the PAM, which has been termed the seed sequence, is of particular importance as mismatches between the target and this part of the crRNA will abrogate interference [40, 41]. Finally the target DNA is broken in ways that depend on each specific type of CRISPR-Cas system. In type I systems the Cas3 nuclease will progressively degrade the target strand thanks to its helicase and exonuclease activity [15]. In type II systems a blunt double strand break is created [8, 16, 42], while in type V systems a staggered cut is produced [17]. Combined with the action of host nucleases this results in the destruction of target DNA molecules. In type III systems the complex of Cas proteins binds to messenger RNAs rather than DNA, but this activates both an RNAse and DNAse activity that leads to the destruction of both the target mRNA and DNA. [43, 44]

The reliance on the PAM motif enables self vs. non-self discrimination, as the CRISPR array does not carry a PAM motif next to the spacer, and therefore will not damage its own DNA. Type III systems do not rely on a PAM motif for this purpose but rather on the complementarity between the crRNA handle and the sequence just upstream [45].

1.3 – Methods to study CRISPR-Cas activity

The experimental procedures described in this chapter focus on the use of type II-A CRISPR systems in *Staphylococcus aureus*, but the general concepts can be applied to other experimental systems. A previous volume of Methods in Molecular Biology has been released

focusing on CRISPR. The book, “CRISPR, Methods and Protocols” can offer more insight on working with CRISPR for the first time. We particularly recommend chapter 10 (on interference) and chapter 13 (on acquisition), which include protocols that can be useful in the context of this chapter [46, 47]. A list of references of experimental work carried on various types of CRISPR-Cas systems is provided in table 1.

Studying the activity of a CRISPR system involves different methods for each step of CRISPR immunity: adaptation, expression and interference. This chapter will provide methods to investigate adaptation and interference. Methods used to study expression involve the detection and characterization of small RNA molecules and their processing. We refer the reader to the following article for protocols dedicated to these aspects [48], as well as to two chapters from the “CRISPR, Methods and Protocols” book on the same topics [49, 50].

Table 1 = List of references of experimental work carried on various types of CRISPR-Cas systems

<table>
<tr><th>Subject</th><th>Type</th><th>Adaptation</th><th>Interference</th></tr>
<tr><td>General</td><td></td><td>Amitai and Sorek. (2016) [56]
Hynes et al. (2017) [57]
Sternberg et al. (2016) [59]
Wright et al. (2016) [61]</td><td>Barrangou et al. (2007) [21]
Barrangou. (2013) [58]
Marraffini and Sontheimer. (2010) [60]
Marraffini. (2015) [39]
Wiedenheft et al. (2012) [62]</td></tr>
<tr><td rowspan="5">Type I CRISPR-Cas systems</td><td colspan="2">Brouns et al. (2008) [34]
Jore et al. (2011) [63]
Marraffini and Sontheimer. (2009) [64]
Van der Oost et al. (2009) [65]</td><td></td></tr>
<tr><td>Type I-A</td><td>Cady et al. (2012) [66]</td><td></td></tr>
<tr><td>Type I-B</td><td>Datsenko et al. (2012) [31]
Diez-Villaseñor et al. (2013) [67]</td><td rowspan="2">n/a</td></tr>
<tr><td>Type I-E</td><td>Arslan et al. (2014) [25]
Erdmann and Garret. (2012) [68]
Erdmann et al. (2014) [69]
Levy et al. (2015) [70]
Li et al. (2014) [71]
Li et al. (2014) [72]
Savitskaya et al. (2013) [73]
Shmakov et al. (2014) [74]</td></tr>
<tr><td>Type I-F</td><td>Richter et al. (2014) [75]
Yosef et al. (2012) [77]
Yosef et al. (2013) [78]</td><td>Dwarakanat et al. (2015) [76]</td></tr>
</table>

Type II CRISPR-Cas systems	Type II-A	Barrangou et al. (2007) [21] Deveau et al. (2008) [80] Garneau et al. (2010) [8] Heler et al. (2015) [27] Wei et al. (2015) [81]	Chylinski et al. (2013) [79] Heler et al. (2015) [27] Jinek et al. (2012) [16] Sternberg et al. (2014) [38]
Type III CRISPR-Cas systems	Type III	Hale et al. (2012) [82] Samai et al. (2015) [43]	
	Type III-A	n/a	Goldberg et al. (2014) [54] Marraffini and Sontheimeier. (2008) [7] Staals et al. (2014) [83] Tamulaitis et al. (2014) [84]
	Type III-B	Erdmann and Garret. (2012) [85] Erdmann et al. (2014) [69]	Deng et al. (2013) [86] Peng et al. (2015) [87]
Other types of CRISPR-Cas systems	Type V	n/a	Gao et al. (2016) [88] Sontheimer and Wolfe. (2015) [89] Zetsche et al. (2015) [90]
	Type VI	n/a	Abudayyeh et al. (2016) [19] Shmakov et al. (2015) [91]

The study of interference requires a CRISPR array containing a spacer targeting an exogenous sequence, which could either be carried by a phage or a plasmid. Interference can only be observed if a PAM motif is present next to the target sequence. If the PAM is not known, one can try to identify proper targets by blasting the CRISPR spacers against sequence databases in the hope to find natural targets, which can then be cloned. Ideally knowledge of the PAM should be obtained and is a prerequisite for any serious investigation of interference. Methods to characterise PAM sequences have been described elsewhere [51, 30]. With knowledge of the PAM, one can easily modify a plasmid to clone a target that will match one of the spacers already present in the CRISPR array. When studying CRISPR interference against phages, it is frequently easier to modify the CRISPR to target the phage rather than modifying the phage. CRISPR arrays can be provided on a plasmid, and modified with methods inspired by golden-gate assembly for the easy cloning of novel spacers. The goal is then to investigate whether target genetic elements are blocked by the CRISPR system (no transformation, or no phage plaques), while non-target sequences can readily enter the cell. An alternative method to quickly check the activity of a CRISPR system is to clone a CRISPR array on a plasmid and program it to target a sequence in the bacteria's own chromosome. Cleavage by Cas nucleases in the chromosome leads to cell death [9, 52]. If the CRISPR system is active, then such a CRISPR plasmid cannot be transformed in the recipient cells, while a control non-targeting CRISPR plasmid can.

To determine adaptation activity, one simply needs to challenge a bacterium with phages and monitor the addition of spacers in the CRISPR array. Bacteria can survive the infection using several mechanisms (reviewed in ref [53]), besides spacer acquisition in the CRISPR array. During an adaptation experiment, a mixture of bacteria that resist through different mechanisms is thus frequently obtained. Checking for an adaptation event is commonly done by PCR, as described below. Primers are designed so that the PCR product will be longer if novel spacers are incorporated. This can simply be visualised on an agarose gel, and the product can also be sequenced. High-throughput sequencing can also be used to determine the general profile of acquired sequences. When performing adaptation experiments, also remember that primed adaptation occurs at much higher frequencies than naïve adaptation. Modifying the CRISPR array or the phage sequence so that a mismatched target is present can greatly facilitate measurements of spacer acquisition. In the absence of phage infection assays, it might still be possible to assess adaptation activity. Several studies have focused on the acquisition of spacers from protospacers carried on plasmids or in the bacteria's own chromosome. Such events will lead to plasmid loss or cell death and are harder to capture. A PCR trick using a primer with a mismatched 3' end can be used to amplify rare acquisition events and analyse them.

The vast majority of *Staphylococcus aureus* isolates do not carry a CRISPR-Cas system. However, the group of Luciano Marraffini has been able to transfer both type III and type II CRISPR systems in this organism and established it as a model organism to study CRISPR [27, 54]. Here we will describe methods using the type II-A CRISPR-cas system from *Streptococcus pyogenes* cloned on a staphylococcal vector. In particular, plasmid pDB114 carries the tracrRNA, Cas9 and a minimal CRISPR array with a single spacer displaying BsaI restriction sites for the easy cloning of new spacers [55]. It is used in interference assays. Plasmid pRH087 carries the complete CRISPR-Cas operon, and the first repeat-spacer-repeat from *S. pyogenes*, and is used in adaptation assays [27].

2. Materials

2.1 – Working with *S. aureus*

1. *S. aureus* is grown in tryptic soy broth (TSB). Use ready-to-use TSB powdered medium to prepare TSB in a liquid solution, sterilise it using an autoclave and store at room temperature until use.
2. Tryptic soy agar (TSA) is TSB supplemented with 1% agar. Prepare TSA the same way as TSB, adding a final concentration of agarose 1% before sterilization.
3. Making *S. aureus* electro-competent cells requires ice-cold ddH_2O and an ice-cold 10% glycerol solution. Both should be autoclaved or filtered before use.
4. A lysis procedure is required to perform PCR on colonies. This requires lysostaphin and the following lysis buffer: 250mM KCl, 5mM $MgCl_2$, 50mM Tris-HCl at pH 9.0, 0.5% Triton X-100.

2.2 – Working with phage ϕNM4

1. When working with ϕNM4 it is important to remember to add $CaCl_2$ to a final concentration of 5mM to the growth media.
2. A layer of soft top-agar is used in many phage protocols. TSA top-agar is TSA with an agar concentration of 0.7%.

2.3 – List of primers for testing spacer acquisition in plasmid pRH087

Table 2 : List of primers used throughout this chapter

Oligo 1 (on-plate method)	5'-CGAAATTTTTTAGACAAAAATAGTC-3'
Oligo 2 (on-plate method)	5'-AAAACAAAAAGCGCAAGAAGAAATCAACCAGCGCA-3'
Oligo 3 (in-liquid method)	5'-GGCTTTTCAAGACTGAAGTCTAG-3'
Oligo 4 (in-liquid method)	5'-AAAACAGCATAGCTCTAAAACG-3',
Oligo 5 (in-liquid method)	5'- AAAACAGCATAGCTCTAAAACA-3'
Oligo 6 (in-liquid method)	5'-AAAACAGCATAGCTCTAAAACT-3'

3. Methods

3.1 – Programming the system

3.1.1 – Reprogramming the CRISPR array

Protospacer selection is guided by the presence of a PAM next to the target sequence. The *S. pyogenes* Cas9 protein (SpCas9) recognises a 5'-NGG-3' PAM motif 3' of the target. It is involved in both protospacer selection and target recognition. The 30 nucleotides upstream of the PAM sequence can be cloned into the CRISPR array to make a spacer. The protospacer should be chosen as depicted in figure 1A and cloned into the CRISPR array as depicted in figure 1B.

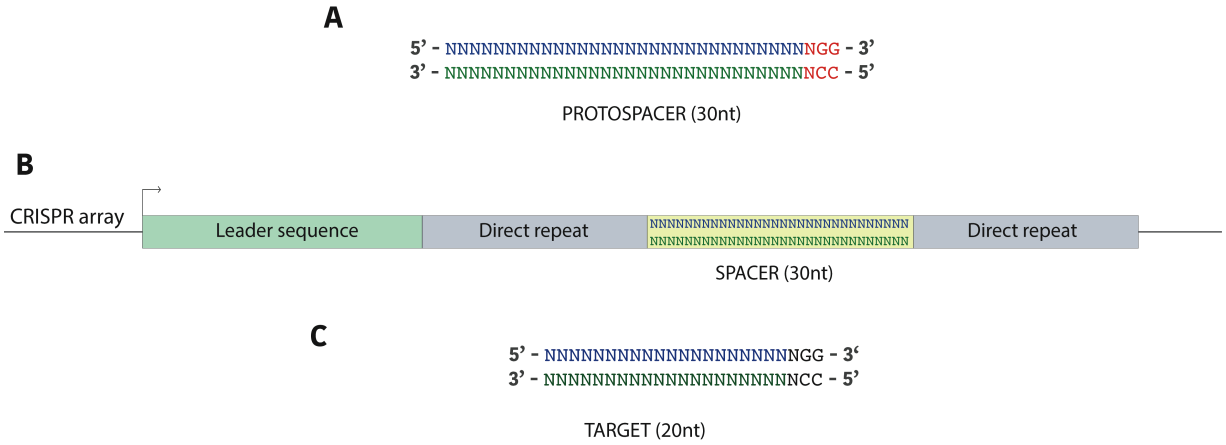


Figure 1 – Protospacer selection and target design

(A) Protospacer selection is guided by the presence of a PAM next to the target sequence. In the case of SpCas9, the NGG PAM should be located immediately downstream from the target sequence. The protospacer consists in the 30nt upstream the PAM sequence. (B) After being cloned into the CRISPR array, the spacer sequence is inserted between direct repeats and matches the 30nt of the protospacer. (C) Although spacers have a length of 30nt, after processing of the pre-crRNA, crRNA only carry 20nt of homology to the target. Therefore, when performing interference experiments only the last 20nt of the spacer followed by a proper PAM have to be cloned in the target vector.

Spacers can be cloned in vector pDB114 or equivalent following the protocol below. On plasmid pDB114, the CRISPR array has been modified to contain a single spacer with two *Bsa*I restriction sites, cutting just outside of the spacer in the repeats. This enables a scar-less replacement of the spacer by any novel sequence of choice. Two oligonucleotides following the design described in figure 2 can be synthesised, phosphorylated and annealed.

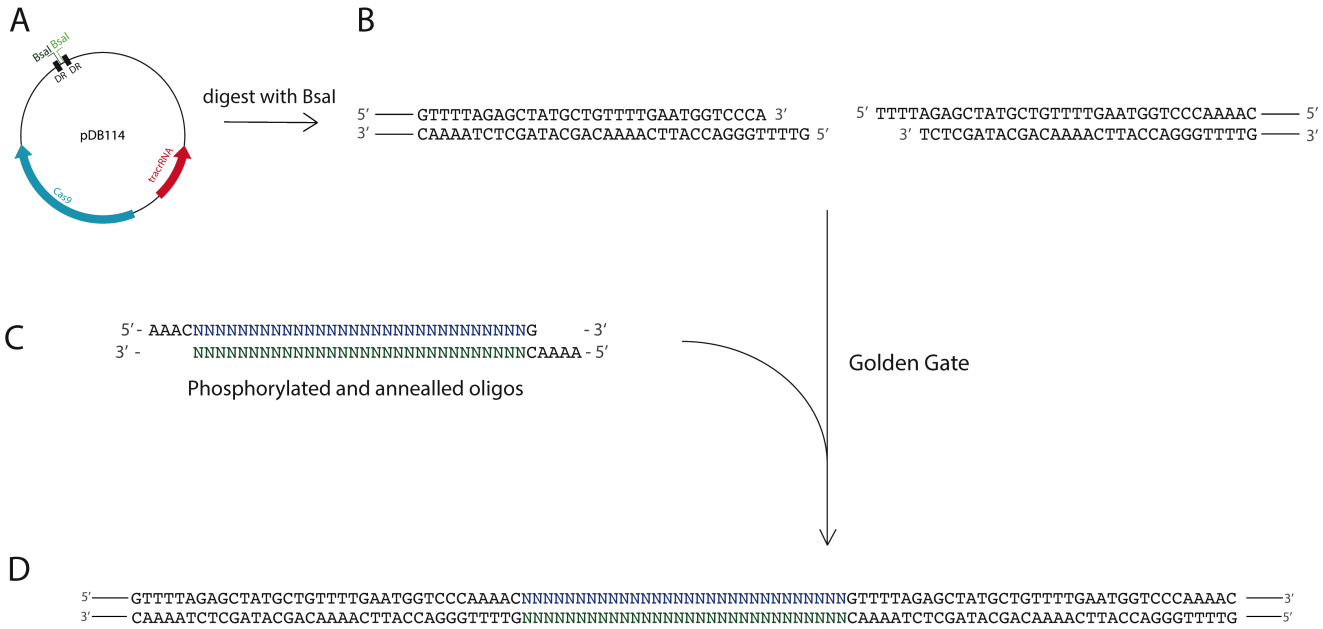


Figure 2 – Golden gate strategy for reprogramming the CRISPR array

(A) Vector pDB114 has two *Bsa*I restriction sites that cut just outside of the spacer within the repeats. (B) Digesting pDB114 with *Bsa*I gives place to sticky ends that can then be complemented by a correctly designed spacer. (C) Two oligos with the appropriate target sequence have to be designed, adding the overhangs indicated in the figure. After phosphorylation and annealing, (D) the oligos can be cloned in plasmid pDB114 through Golden Gate assembly.

Oligonucleotide phosphorylation and annealing:

1. Setup the following mix: 15µL of each of the 10µM dilution of both oligos, and a mix of T4 PNK, its appropriate buffer and H_2O to a final volume of 50µL.
2. Incubate the mix at 37ºC for 30 min.
3. Add 2.5µL of NaCl at 1M. (Salts help annealing)
4. Incubate the mix at 95ºC for 5 minutes in a heat block and slowly cool the mix down to room temperature by leaving the block on the bench until the temperature decreases to bellow 40ºC.

The annealed oligos can then be cloned following the Golden Gate cloning technique:

1. In a PCR tube, mix 100ng of the plasmid miniprep (*see* **Note 1**), 2µL of the annealed oligos (1/10 dilution of the protocol above), and then 20U of *Bsa*I, 400 cohesive-ends U of T4 ligase, the appropriate buffer (with ATP) and H_2O to a final volume of 10µL.
2. In a thermocycler set the following cycle: 25 cyles of 3 min at 37ºC and 4 min at 16ºC; 1 cycle of 5 min at 50ºC and 5 min at 80ºC.
3. Dialyze the mix, and introduce in *S. aureus* RN4220 by electroporation (2500V, 25µF, 100 Ω and 2mm cuvettes).

3.1.2 –Constructing a plasmid that is targeted by the CRISPR-Cas system

Reprogramming a CRISPR array as described above requires a plasmid carrying a CRISPR array already modified to contain a spacer with restriction sites. Rather than constructing such a plasmid, it can be easier to measure the interference activity of spacers already present in the CRISPR array. In order to identify CRISPR arrays and the spacers they carry, online tools such as CRISPRDetect [11] or CRISPR finder [10] can be used. A target with the proper PAM sequence can then be cloned on a plasmid. In the case of the *S. pyogenes* type II-A CRISPR-systems, spacers have a length of 30nt, but the processed crRNA only carries a 20nt sequence of homology to the target. So it is sufficient to clone the last 20nt of the spacer followed by a proper NGG, as depicted in figure 1C.
This sequence can be introduced at any position and orientation on a plasmid following common molecular cloning techniques.

3.2 – Interference protocols

3.2.1 – Measuring CRISPR Interference using phages

1. Launch an overnight culture of the *S. aureus* strain carrying the re-programmed CRISPR, and as a negative control, of a strain carrying the CRISPR programmed to target some sequence not present in phage ϕNM4.
2. Take 100µL of the ON culture and supplement it with enough $CaCl_2$ to reach a 5mM concentration once the TSA top-agar is added.
3. Add 5mL of TSA top agar to the mix (*see* **Note 2**) and quickly pour it over TSA+$CaCl_2$ plates.
4. When the top agar layer has set, spot serial dilutions of the phage stock over the top agar layer. Perform dilutions down to 10^{-8}. We recommend spotting 2µL of each dilution.
5. Incubate at 37ºC overnight.
6. Next day compare the number of PFUs obtained on the strain with the CRISPR programmed against the phage, to the number of PFUs obtained on the strain with the target-less CRISPR.

3.2.2 – Interference using plasmids

1. Launch an overnight culture of the *S. aureus* strain carrying the CRISPR-Cas system in TSB supplemented with appropriate antibiotic (*see* **Note 3**).
2. Dilute the overnight culture 1:100 into 100mL of fresh TSB supplemented with the appropriate antibiotics, and incubate at 37°C shaking at 250rpm.

3. Wait until the culture reaches an optical density OD_{600nm} of 0.8.
4. Chill the cells on ice for 10 min, then centrifuge at 4000g for 10min.
5. Wash the cells twice with 20mL of ice-cold water, and once with 10mL of ice-cold glycerol 10%.
6. Re-suspend the cells in 1mL of 10% glycerol and stock at -80ºC. Cells are now electro-competent.
7. Electroporate up to 5µL of a miniprep of the plasmid carrying the CRISPR target (*see* **Notes 4 and 5**). Recover the cells in 1mL of TSB, at 37ºC with shaking.
8. Plate the cells on TSA supplemented with the appropriate antibiotics.

Next day, compare the number of CFUs obtained between conditions where the CRISPR system targets the plasmid and condition where it does not.

3.3 – Adaptation protocols – spacer acquisition assays

3.3.1 – Adaptation using phages

3.3.1.1 - Spacer acquisition assay – on-plate method

1. Launch an overnight culture of a *S. aureus* strain that carries the CRISPR-Cas9 system (*see* **Note 6**). Overnight cultures should be launched on fresh TSB supplemented with appropriate antibiotic (*see* **Note 3**).
2. Next day, mix 100µL of cells from the overnight culture with ϕNM4 at an m.o.i. (*see* **Note 7**) of 1 in 5mL of top agar (*see* **Note 2**) supplemented with the appropriate antibiotic and 5mM $CaCl_2$ (*see* **Note 8**).
3. Quickly (top agar sets very fast) pour the mixture on top of TSA plates supplemented with the appropriate antibiotic and incubate at 37ºC overnight.
4. Colonies observed in the plate correspond to cells that survived phage infection. Restreak isolated colonies in TSA plates supplemented with the appropriate antibiotic (*see* **Notes 3 and 9**). Incubate the plates at 37ºC overnight.
5. To check for spacer acquisition, pick individual colonies (*see* **Note 10**) and resuspend them in lysis buffer (250mM KCl, 5mM $MgCl_2$, 50mM Tris-HCl at pH 9.0, 0.5% Triton X-100) (*see* **Note 11**) with 50ng μL^{-1} lysostaphin (*see* **Note 12**).
6. Incubate the samples at 37ºC for 10 min and then at 98ºC for 10 min as well (*see* **Note 13**).
7. Centrifuge the samples at 11000g for 1 min.
8. Use between 1µL of the supernatant of each sample in a PCR reaction using as primers the following oligos: oligo 1 and oligo 2 (*see* **Note 14**).

9. Analyse the PCR reactions on 2% agarose gels. In the case of adaptation not having occurred, the size of the amplification will be 100bp. In case of happening of an adaptation event, the size of the amplification will be 100bp+66bp, since a new spacer-repeat will have been added to the array (*see* **Note 15**).

3.3.1.2 - Spacer acquisition assay – in-liquid method

In this assay, rather than looking for adaptation events in single colonies, we amplify adapted CRISPR arrays in the bulk culture. For this purpose we use a mixture of primers carrying mismatched 3'ends that will preferentially amplify CRISPR arrays that have captured a novel spacer, even if they represent only a small fraction of the CRISPR arrays present in the culture (*see* **Note 16**).

1. Launch an overnight culture of a *S. aureus* strain that carries the CRISPR-Cas9 system. Overnight cultures should be launched on fresh TSB supplemented with appropriate antibiotic (*see* **Note 3**).
2. Dilute the overnight culture 1:100 into 10mL of fresh TSB supplemented with the appropriate antibiotics and 5mM $CaCl_2$.
3. Wait until the culture reaches an optic density OD_{600nm} of 0.4 and infect with ϕNM4 to an m.o.i. (*see* **Note 7**) of 1.
4. Incubate at 37ºC for 16h (*see* **Note 17**).
5. Perform plasmid extraction of the cultures using any desired plasmid extraction kit. It is important nonetheless to modify the plasmid extraction protocol slightly. On the step of resuspending the cells after the first centrifugation, add lysostaphin (*see* **Note 12**) to a final concentration of 50ng μL^{-1} and incubate at 37ºC for 1 hour. Then continue with the standard protocol.
6. Use 100ng of plasmid DNA to amplify the CRISPR *locus*. As primers use a mix consisting of 3 parts of oligo 3 and 1 part each of oligos 4, 5 and 6 (*see* **Note 16**).
7. Analyse the PCR reactions on 2% agarose gels. In case of no adaptation only one band will be amplified. In case of adaptation, two bands will be present. Analysis of the bands' strength allows for quantification of the adaptation.

4. Notes

1. Although 100ng is the recommended amount of plasmid to use, the protocol works fine with smaller amounts. A range of 50-150ng of plasmid can be used.
2. TSA top agar is a 0.7% agar version of TSA. The lower agar concentration allows phage diffusion and better plaque formation.
3. Working concentrations of antibiotics for *S. aureus* are as follows: chloramphenicol 10µg mL^{-1}, kanamycin 30µg mL^{-1}, erythromycin 10µg mL^{-1} and tetracyclin 5µg mL^{-1}.
4. The electroporation setup that we normally use is as follows: 2mm electroporation cuvettes, 50µL of electrocompetent cells, 2500V, 25µF and 100Ω.
5. We recommend using as a negative control the exact same plasmid but without a target, or with a mutation in the PAM sequence.
6. In this assay the CRISPR array should not contain a spacer targeting the phage with a perfect match. Imperfect targets can be used to increase adaptation frequency through primed adaptation.
7. Multiplicity of infection is the ratio of viral particles to the number of target cells.
8. Adding $CaCl_2$ is important because many phages require Ca^{++} in solution to form plaques. The cation is needed for nucleic acid injection and/or efficient adsorption to the cell wall binding sites.
9. Colonies obtained at this step can contain mixtures of adapted and non-adapted cells, which can also still undergo attack by phages and mutant phages. Restreaking enables to obtain a pure colony that is easier to analyse.
10. When picking the individual colonies there is no need to pick a lot of cells. Just lightly touching the colony should be enough.
11. We recommend resuspending each colony in 20-40µL of lysis buffer.
12. Lysostaphin is a metalloendopeptidase that is able to cleave crosslinking bridges of pentaglycine found in the peptidoglycan layer of *S. aureus.* Using it in this step is needed to break *S. aureus'* wall and therefore retrieve the DNA in solution.
13. We recommend performing all these steps in PCR strips. This cycle of temperatures can be easily set-up in a thermocycler.
14. One of these primers anneals with the leader region of the CRISPR array (a sequence upstream the CRISR array) and the other anneals with the first repeat.
15. Fragments need to be separated in 2% agarose gels. This concentration of agarose is needed to nicely separate the fragments to be detected.
16. The first primer anneals on the leader region of the CRISPR, the other three anneal on the repeat and differ only in their 3' end nucleotide. This last nucleotide does not match the leader sequence. Only upon spacer acquisition will the 3' end of these oligonucleotides anneal

to the template DNA, enabling efficient amplification. For more information on this technique read the supplementary information of the following article [27].

17. The duration of the incubation can be adjusted depending on the purpose of the experiment. Longer incubations can lead to the amplification of cells that first captured spacers. Spacers might also be captured from mutant phages that can arise during the course of the incubation.

References

1. Makarova KS, Wolf YI, Alkhnbashi OS, et al (2015) An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13:722–736. doi: 10.1038/nrmicro3569
2. Ishino Y, Shinagawa H, Makino K, et al (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol 169:5429–33. doi: 10.1128/jb.169.12.5429-5433.1987
3. Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151:2551–2561. doi: 10.1099/mic.0.28048-0
4. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60:174–182. doi: 10.1007/s00239-004-0046-3
5. Pourcel C, Salvignol G, Vergnaud G (2005) CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151:653–663. doi: 10.1099/mic.0.27437-0
6. Barrangou R, Fremaux C, Deveau H, et al (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science (80-) 315:1709 LP-1712.
7. Marraffini LA, Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science (80-) 322:1843–1845. doi: 10.1126/science.1165771
8. Garneau JE, Dupuis ME, Villion M, et al (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468:67–71. doi: 10.1038/nature09523
9. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA (2012) CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. Cell Host Microbe 12:177–186. doi: 10.1016/j.chom.2012.06.003
10. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. doi: 10.1093/nar/gkm360
11. Biswas A, Staals RHJ, Morales SE, et al (2016) CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 17:356. doi: 10.1186/s12864-016-2627-0
12. Shmakov S, Smargon A, Scott D, et al (2017) Diversity and evolution of class 2 CRISPR–Cas systems. Nat Rev Microbiol 15:169–182. doi: 10.1038/nrmicro.2016.184
13. Sinkunas T, Gasiunas G, Fremaux C, et al (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J 30:1335–1342. doi: 10.1038/emboj.2011.41
14. Gong B, Shin M, Sun J, et al (2014) Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proc Natl Acad Sci U S A 111:16359–16364. doi: 10.1073/pnas.1410806111
15. Huo Y, Nam KH, Ding F, et al (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol 21:771–777. doi: 10.1038/nsmb.2875
16. Jinek M, Chylinski K, Fonfara I, et al (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science (80-) 337:816–822. doi: 10.1126/science.1225829
17. Zetsche B, Gootenberg JS, Abudayyeh OO, et al (2015) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell 163:759–771. doi: 10.1016/j.cell.2015.09.038
18. Shmakov S, Abudayyeh OO, Makarova KS, et al (2015) Discovery and Functional Characterization

of Diverse Class 2 CRISPR-Cas Systems. Mol Cell 60:385–397. doi: 10.1016/j.molcel.2015.10.008
19. Abudayyeh OO, Gootenberg JS, Konermann S, et al (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science 353:aaf5573. doi: 10.1126/science.aaf5573
20. Modell JW, Jiang W, Marraffini LA (2017) CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature 544:101–104.
21. Barrangou R, Fremaux C, Deveau H, et al (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-). doi: 10.1126/science.1138140
22. Wiedenheft B, Zhou K, Jinek M, et al (2009) Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. Structure 17:904–912. doi: 10.1016/j.str.2009.03.019
23. Beloglazova N, Brown G, Zimmerman MD, et al (2008) A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J Biol Chem 283:20361–20371. doi: 10.1074/jbc.M803225200
24. Yosef I, Goren MG, Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res 40:5569–5576. doi: 10.1093/nar/gks216
25. Arslan Z, Hermanns V, Wurm R, et al (2014) Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. Nucleic Acids Res 42:7884–7893. doi: 10.1093/nar/gku510
26. Nuñez JK, Kranzusch PJ, Noeske J, et al (2014) Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nat Struct Mol Biol 21:528–534.
27. Heler R, Samai P, Modell JW, et al (2015) Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature 519:199–202.
28. Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ (2012) CRISPR interference directs strand specific spacer acquisition. PLoS One 7:1–7. doi: 10.1371/journal.pone.0035888
29. Leenay RT, Maksimchuk KR, Slotkowski RA, et al (2015) Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell 1–11. doi: 10.1016/j.molcel.2016.02.031
30. Karvelis T, Gasiunas G, Siksnys V (2017) Methods for decoding Cas9 protospacer adjacent motif (PAM) sequences: A brief overview. Methods. doi: 10.1016/j.ymeth.2017.03.006
31. Datsenko K a., Pougach K, Tikhonov A, et al (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun 3:945. doi: 10.1038/ncomms1937
32. Fineran PC, Charpentier E (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. Virology 434:202–209. doi: 10.1016/j.virol.2012.10.003
33. Semenova E, Savitskaya E, Musharova O, et al (2016) Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. Proc Natl Acad Sci 113:7626–7631. doi: 10.1073/pnas.1602639113
34. Brouns SJJ, Jore MM, Lundgren M, et al (2008) Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science (80-) 321:960 LP-964.
35. Deltcheva E, Chylinski K, Sharma CM, et al (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471:602–607.
36. Haurwitz RE, Jinek M, Wiedenheft B, et al (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 329:1355–1358. doi: 10.1126/science.1192272
37. Carte J, Wang R, Li H, et al (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 22:3489–3496. doi: 10.1101/gad.1742908
38. Sternberg SH, Redding S, Jinek M, et al (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507:62–67. doi: 10.1038/nature13011
39. Marraffini LA (2015) CRISPR-Cas immunity in prokaryotes. Nature 526:55–61. doi:

10.1038/nature15386
40. Semenova E, Jore MM, Datsenko K a, et al (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A 108:10098–10103. doi: 10.1073/pnas.1104144108
41. Wiedenheft B, van Duijn E, Bultema JB, et al (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci 108:10092–10097. doi: 10.1073/pnas.1102716108
42. Gasiunas G, Barrangou R, Horvath P, Siksnys V (2012) Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci 109:E2579–E2586. doi: 10.1073/pnas.1208507109
43. Samai P, Pyenson N, Jiang W, et al (2015) Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. Cell 161:1164–1174. doi: 10.1016/j.cell.2015.04.027
44. Jiang W, Samai P, Marraffini LA (2016) Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell 164:710–721. doi: 10.1016/j.cell.2015.12.053
45. Marraffini LA, Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463:568–571.
46. Dupuis M-ève, Barrangou R, Moineau S (2015) Procedures for Generating CRISPR Mutants with Novel Spacers Acquired from Viruses or Plasmids. doi: 10.1007/978-1-4939-2687-9
47. Almendros C, Mojica FJM (2015) Exploring CRISPR Interference by Transformation with Plasmid Mixtures: Identification of Target Interference Motifs in Escherichia coli. Methods Mol Biol 1311:161–170. doi: 10.1007/978-1-4939-2687-9_10
48. Heidrich N, Dugar G, Vogel J, Sharma CM (2015) Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. Methods Mol Biol 1311:1–21. doi: 10.1007/978-1-4939-2687-9_1
49. Waghmare SP, Nwokeoji AO, Dickman MJ (2015) Analysis of crRNA Using Liquid Chromatography Electrospray Ionization Mass Spectrometry (LC ESI MS). Methods Mol Biol 1311:133–145. doi: 10.1007/978-1-4939-2687-9_8
50. Garside EL, MacMillan AM (2015) Analysis of CRISPR Pre-crRNA Cleavage. Methods Mol Biol 1311:35–46. doi: 10.1007/978-1-4939-2687-9_3
51. Leenay RT, Beisel CL (2016) Deciphering, communicating, and engineering the CRISPR PAM. J Mol Biol 429:177–191. doi: 10.1016/j.jmb.2016.11.024
52. Edgar R, Qimron U (2010) The Escherichia coli CRISPR system protects from ?? lysogenization, lysogens, and prophage induction. J Bacteriol 192:6291–6294. doi: 10.1128/JB.00644-10
53. Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Micro 8:317–327.
54. Goldberg GW, Jiang W, Bikard D, Marraffini L a (2014) Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature 514:633–637. doi: 10.1038/nature13637
55. Bikard D, Euler CW, Jiang W, et al (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotech 32:1146–1150.
56. Amitai G, Sorek R (2016) CRISPR–Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol advance on:67–76. doi: 10.1038/nrmicro.2015.14
57. Hynes AP, Lemay M, Trudel L, et al (2017) Detecting natural adaptation of the Streptococcus thermophilus CRISPR-Cas systems in research and classroom settings. Nat Protoc 12:547–565. doi: 10.1038/nprot.2016.186
58. Barrangou R (2013) CRISPR-Cas systems and RNA-guided interference. Wiley Interdiscip Rev RNA 4:267–278. doi: 10.1002/wrna.1159
59. Sternberg SH, Richter H, Charpentier E, Qimron U (2016) Adaptation in CRISPR-Cas Systems.

Mol Cell 61:797–808. doi: 10.1016/j.molcel.2016.01.030
60. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11:181–190. doi: 10.1038/nrg2749
61. Wright A V., Nu??ez JK, Doudna JA (2016) Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell 164:29–44. doi: 10.1016/j.cell.2015.12.035
62. Wiedenheft B, Sternberg SH, Doudna J a. (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature 482:331–338. doi: 10.1038/nature10886
63. Jore MM, Lundgren M, van Duijn E, et al (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol 18:529–536.
64. Marraffini LA, Sontheimer EJ (2009) Invasive DNA, Chopped and in the CRISPR. Structure 17:786–788. doi: 10.1016/j.str.2009.05.002
65. van der Oost J, Jore MM, Westra ER, et al (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34:401–407. doi: 10.1016/j.tibs.2009.05.002
66. Cady KC, Bondy-Denomy J, Heussler GE, et al (2012) The CRISPR/Cas Adaptive Immune System of Pseudomonas aeruginosa Mediates Resistance to Naturally Occurring and Engineered Phages. J Bacteriol 194:5728–5738. doi: 10.1128/JB.01184-12
67. Díez-Villaseñor C, Guzmán NM, Almendros C, et al (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol 10:792–802. doi: 10.4161/rna.24023
68. Erdmann S, Garrett RA (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol. doi: 10.1111/j.1365-2958.2012.08171.x
69. Erdmann S, Le Moine Bauer S, Garrett RA (2014) Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. Mol Microbiol 91:900–917. doi: 10.1111/mmi.12503
70. Levy A, Goren MG, Yosef I, et al (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520:505–510.
71. Li M, Wang R, Zhao D, Xiang H (2014) Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res 42:2483–2492. doi: 10.1093/nar/gkt1154
72. Li M, Wang R, Xiang H (2014) Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. Nucleic Acids Res 42:7226–7235. doi: 10.1093/nar/gku389
73. Savitskaya E, Semenova E, Dedkov V, et al (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biol 10:716–25. doi: 10.4161/rna.24325
74. Shmakov S, Savitskaya E, Semenova E, et al (2014) Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res 42:5907–5916. doi: 10.1093/nar/gku226
75. Richter C, Dy RL, McKenzie RE, et al (2014) Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res 42:8516–8526. doi: 10.1093/nar/gku527
76. Dwarakanath S, Brenzinger S, Gleditzsch D, et al (2015) Interference activity of a minimal Type I CRISPR-Cas system from Shewanella putrefaciens. Nucleic Acids Res 43:8913–8923. doi: 10.1093/nar/gkv882
77. Yosef I, Goren MG, Qimron U (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. doi: 10.1093/nar/gks216
78. Yosef I, Shitrit D, Goren MG, et al (2013) DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proc Natl Acad Sci U S A 110:14396–401. doi: 10.1073/pnas.1300108110

79. Chylinski K, Le Rhun A, Charpentier E (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol 10:726–37. doi: 10.4161/rna.24321
80. Deveau H, Barrangou R, Garneau JE, et al (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol 190:1390–1400. doi: 10.1128/JB.01412-07
81. Wei Y, Chesne MT, Terns RM, Terns MP (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Res 43:1749–1758. doi: 10.1093/nar/gku1407
82. Hale CR, Majumdar S, Elmore J, et al (2012) Essential Features and Rational Design of CRISPR RNAs that Function with the Cas RAMP Module Complex to Cleave RNAs. Mol Cell 45:292–302. doi: 10.1016/j.molcel.2011.10.023
83. Staals RHJ, Zhu Y, Taylor DW, et al (2014) RNA Targeting by the Type III-A CRISPR-Cas Csm Complex of Thermus thermophilus. Mol Cell 56:518–530. doi: 10.1016/j.molcel.2014.10.005
84. Tamulaitis G, Kazlauskiene M, Manakova E, et al (2014) Programmable RNA Shredding by the Type III-A CRISPR-Cas System of Streptococcus thermophilus. Mol Cell 56:506–517. doi: 10.1016/j.molcel.2014.09.027
85. Erdmann S, Garrett RA (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol 85:1044–1056. doi: 10.1111/j.1365-2958.2012.08171.x
86. Deng L, Garrett RA, Shah SA, et al (2013) A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol 87:1088–1099. doi: 10.1111/mmi.12152
87. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43:1–10. doi: 10.1093/nar/gkv585
88. Gao M, Monian P, Pan Q, et al (2016) Ferroptosis is an autophagic cell death process. Cell Res 26:1021–32. doi: 10.1038/cr.2016.95
89. Sontheimer EJ, Wolfe SA (2015) Cas9 gets a classmate. Nat Biotech 33:1240–1241.
90. Zetsche B, Gootenberg JS, Abudayyeh OO, et al (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163:759–771. doi: 10.1016/j.cell.2015.09.038
91. Shmakov S, Abudayyeh OO, Makarova KS, et al (2015) Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell 60:385–397. doi: 10.1016/j.molcel.2015.10.008

Chapter 3

Distribution, organization, interactions and transfer of CRISPR-Cas systems

The work presented in this chapter is still ongoing. Some of the results are preliminary and might be subject to changes.

Aude Bernheim[1,2,3,4], David Bikard[3], Marie Touchon[1,2], Eduardo P.C. Rocha[1,2,*]

[1]Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue Dr Roux, Paris, 75015, France
[2]CNRS, UMR3525, 25-28 rue Dr. Roux, Paris, 75015, France
[3]Synthetic Biology Group, Institut Pasteur, 25-28 rue Dr. Roux, Paris, 75015, France
[4]AgroParisTech, F-75005 Paris, France
[*] Provisional order of the authors

Abstract

CRISPR-Cas are bacteria's adaptive immune system. They are extremely diverse and recent studies have classed them in numerous types and subtypes. Yet, the taxonomic distribution, genetic organization and abundance on mobile genetic elements of these systems has not been assessed. Furthermore, the associations between different CRISPR arrays and Cas systems have not been studied in detail. Here, we detected and analysed in an integrated manner the Cas clusters and CRISPR arrays of 5563 genomes. These systems present very diverse genetic organizations between (multiple) close co-occurring Cas systems and CRISPR arrays. Their analysis reveals frequent co-occurrence between certain Cas systems and, more rarely, some pairs of systems that seem to be counter-selected. This

results in frequent presence of long and complex Cas loci that have components of different types, often in multiple copies, that cannot be easily typed in existing classes. Using logistic regression, we develop a method to identify the Cas types associated with CRISPR arrays and used it to classify most arrays that are not encoded close to Cas loci. This method also allowed to show that plasmids tend to encode CRISPR arrays compatible with the Cas systems encoded by their host genomes. This integrative study quantifies many disparate observations in specific genomes and opens the way to the large-scale study of the interactions between CRISPR and Cas in Prokaryotes.

3.1 Introduction

CRISPR-Cas systems are an adaptive immune system of bacteria and archaea [19, 40, 169, 87]. They are composed of a CRISPR array and a cluster of *cas* genes. CRISPR arrays comprise two types of sequences: repeats and spacers. Repeats are short sequences (typically 20-40 bp) identical in a given CRISPR array. They are interspaced by short and diverse spacer sequences (typically 20-40 bp), which often match sequences from mobile genetic elements. Cas genes encode the proteins involved in the three stages of CRISPR-Cas immunity: adaptation, expression and interference [168].

CRISPR-Cas systems are present in less than half of bacteria [161]. They are extremely diverse and have been recently classified hierarchically in two classes, six types (I to VI) and 21 subtypes [161, 140, 163, 181]. This classification is based on the content and architecture of signature proteins of the Cas cluster [161]. The last surveys of CRISPR-Cas systems abundance and diversity among fully sequenced bacterial genomes included 2740 and 2751 genomes [161, 42]. One study focused on the classification of CRISPR-Cas systems and lacked their precise taxonomic distribution [161]. The other study was restricted to systems with Cas1 because it used its phylogeny to type Cas in a previous classification system [42]. CRISPR-Cas systems are frequently transferred [90, 47, 161]. They have been detected on diverse mobile genetic elements (MGE) like plasmids, phages or transposons [235, 177, 99, 236, 212]. However, no exhaustive study quantified their distribution on MGE.

There have been no quantitative studies integrating both CRISPR and Cas, i.e., assessing the co-occurrence of Cas systems and their association with CRISPR arrays, and detailing the genetic architecture of CRISPR-Cas loci. Yet, previous works have suggested that these traits are important. Functional interactions between CRISPR-Cas systems were recently demonstrated for the subtype I-F and III-B systems [248], where the latter can process crRNAs for the former to improve immune defense. Many *cas* genes are found near CRISPR arrays, but distant arrays (i.e., CRISPR arrays without neighboring cas genes) have also been identified [161, 244]. They can be processed by *cas* genes present in other regions of the genome (in trans) [19] or they may represent inactive systems. The prevalence of such distant elements remains to be assessed.

The goal of this study was to perform an integrated analysis of Cas clusters and CRISPR arrays. To do so, we analysed a large set of bacterial genomes, and some of their mobile genetic elements. Then, using information on genetic composition and organization of CRISPR-Cas systems, we analysed preferential associations between specific subtypes of CRISPR-Cas systems and we developped a method to assign subtypes to CRISPR arrays not encoded near Cas clusters.

3.2 Results

Distribution of CRISPR arrays and Cas clusters

We searched for CRISPR arrays and Cas clusters, independently, in 5563 fully sequenced bacterial genomes (Supplementary table 1). The size of CRISPR arrays varies widely, from a minimum of three repeats (minimal detection threshold) to a maximum of 589 in the Proteobacteria *Haliangium ochraceum DSM 14365* (Figure 3.1.C). Most arrays are small, with 19% of all the detected arrays containing 5 repeats or less. We found CRISPR arrays in 47% and Cas clusters in 42% of the bacterial genomes. Type I Cas systems were by far the most frequent (29%) followed by type II (7%) and type III (6%) (Figure 3.1.A). All of them were found in different phyla. The other types were extremely rare - they were found in only 19 (IV), 12 (V) and 29 (VI) genomes and were restricted to few clades (Proteobacteria, Actinobacteria, Bacteroidetes). The relative abundances of types of Cas clusters are relatively similar to ones reported in previous studies [161, 1], even though Class 1 CRISPR-Cas systems represent 82% of the detected clusters in our dataset while previous studies reported 90%. Some subtypes are present in many clades such as I-B, I-C, II-C, III-A, III-B, III-D while some subtypes are clade specific like type I-D in Cyanobacteria, II-A in Firmicutes or II-B in Proteobacteria. (Figure 3.1.A).

We analysed the distribution of CRISPR-Cas systems in function of genome size, which often is correlated with the presence of bacterial defense pathways [250]. As we have previously reported [27], type II CRISPR-Cas systems are only present in small genomes (Figure 3.1.B). In contrast, type IV systems are usually encoded by large genomes ($>$ 5Mb). Type I and type III systems as well as CRISPR arrays distribution does not seem dependent on genome size.

Recent reports suggested that diverse mobile genetic elements (MGE), including phages and plasmids, encode CRISPR-Cas systems [235, 177, 99, 236, 212]. However, we found only one Cas system and five CRISPR in 1943 genomes of phages (Figure 3.1.D). We then analysed the sequences of prophages to search if successful temperate phages were more likely to encode more CRISPR-Cas systems. We only detected two such systems, six Cas clusters and 33 CRISPR arrays on a very large set of 9926 prophages. CRISPR-Cas systems were more common on a set of 4335 plasmids, both the complete systems (112) and the orphan CRISPR arrays (101) (Figure 3.1.D). Subtypes relative abundance is different on plasmids and chromosomes (Supplementary Figure 2.A). No plasmids encode type II-A CRISPR-Cas systems while type IV are encoded almost exclusively on plasmids. The plasmids carrying either elements (CRISPR or Cas) were larger than the other plasmids (Supplementary Figure 2.B). These results suggest that CRISPR-Cas systems are rare in plasmids and almost never found in phages.

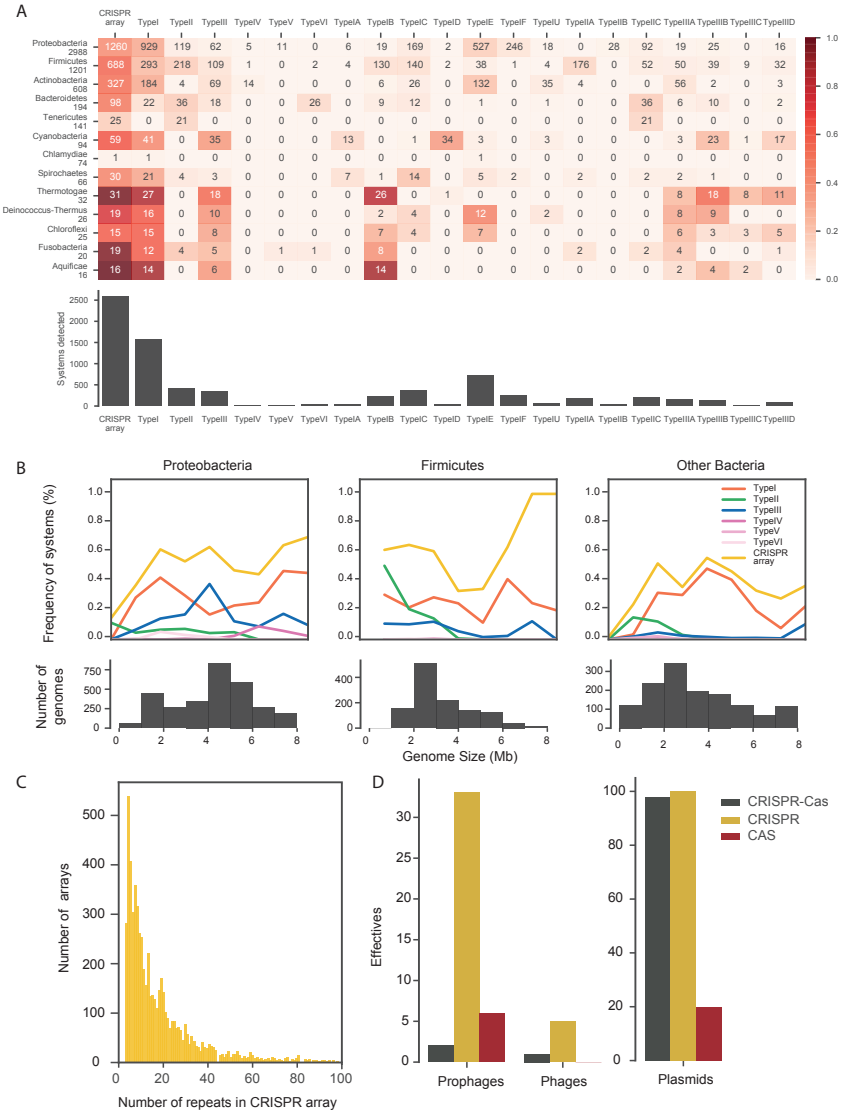


Figure 3.1: Distribution of CRISPR arrays and Cas clusters in bacteria. **A.** Distribution of CRISPR arrays and Cas clusters by clade. The top panel represents the distribution by clades of Cas clusters and CRISPR arrays. Clades are ordered by number of genomes present in the dataset which are indicated on the y axis. Each cell presents the number of systems detected for the specific clade. Each cell is colored proportionally to the frequency of the system in the clade, the darker, the more frequent. The bottom panel is the total number of systems detected in the dataset. **B.** Frequency of CRISPR arrays and Cas clusters in function of genome size. The histogram on the bottom represents the distribution of genome sizes in each clade. The frequency represents the frequency of genomes carrying a system within the genome size range. The three panels represent the two biggest clades in the dataset : Proteobacteria and Firmicutes, and the rest of the dataset. **C.** Histogram of the number of repeats in CRISPR-arrays. **D.** Frequency of CRISPR-Cas systems in prophages, phages and plasmids.

Organization of CRISPR-Cas systems

The canonical view of the organization of a CRISPR-Cas locus is the association of one CRISPR array with one cluster of cas genes. Yet, genomes encode more CRISPR arrays than Cas clusters (Figure 3.2.A). Actually, in genomes encoding several Cas clusters, the number of CRISPR arrays grows faster than the number of Cas clusters (Figure 3.2.A).

We computed the shortest distances between Cas clusters and CRISPR arrays and vice-versa (since there are more CRISPR than Cas, they are not the same). These distributions showed three groups: between 0 and 700bp, between 700bp and 20kb and from 20kb to several millions bp (Supplementary Figure 3.A). The gap between the two first groups may represent non-annotated Cas genes (Supplementary Figure 3.a). Furthermore, some CRISPR-Cas loci encompass several CRISPR-arrays and Cas clusters, in which case the closest Cas from a CRISPR may necessarily be larger than several kb. Hence, we defined a CRISPR-Cas locus as a cluster where Cas and CRISPR were less than 20 kb apart (the probability of the two being encoded less than 20 kb apart randomly in a genome is very low). If several elements (Cas or CRISPR) are clustered together by transitivity, they are part of the same locus.

Based on this threshold, we defined three possible contexts for CRISPR arrays and Cas clusters: the elements are part of a CRISPR-Cas locus if there is at least one Cas and one CRISPR array less than 20kb apart, "orphan" if there is no counterpart element in the genome, and "distant" otherwise. Using this classification, the vast majority of Cas clusters (90%), and a small majority of CRISPR arrays (60%) are part of CRISPR-Cas loci. Around 24% of the latter are distant and 16% are orphans (Figure 3.2.B). Hence, there is an asymmetry in the genetic organization: Cas are much more often systematically associated with CRISPR than the latter with Cas. This tendency remains the same when only considering CRISPR with more than 5 repeats, with respectively 65%, 23%,18% of in loci, distant and orphan CRISPRs.

We then focused specifically on CRISPR-Cas loci and classified them in different architectures based on their number of CRISPR arrays and Cas clusters (Figure 3.2.C). When considering loci encoding two consecutive CRISPR arrays, one possibility is that the arrays actually constitute only one array that was interrupted by a large insertion (e.g., by a transposable element). We classified consecutive CRISPR arrays with identical repeats as one single CRISPR array. They represent 46 loci. Less than half (49%) of the loci encode one single CRISPR array and one single Cas cluster. Many loci have one Cas cluster and two CRISPR arrays (21%), but few have two Cas clusters and only one CRISPR array. Finally, many (30%) loci are more complex than any of the previous organizations, including at least 2 CRISPR arrays and 2 Cas clusters.

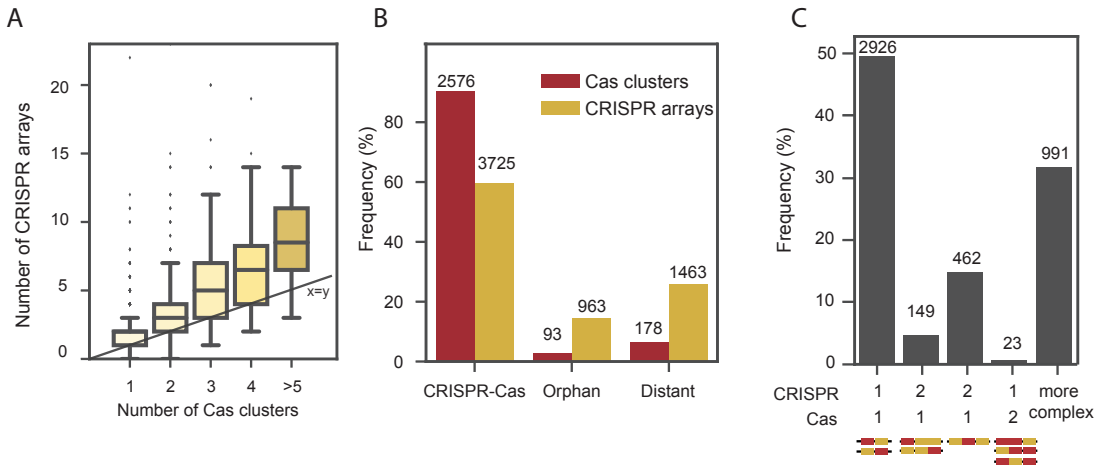


Figure 3.2: Organization of CRISPR-Cas loci.
A. Number of CRISPR arrays in function of number of cas clusters in bacterial genomes (linear regression slope=1.72, 95% confidence interval 1.68,1.76 , P<0.001). **B.** Context of CRISPR-arrays and Cas clusters. **C.** Quantification of the different organizations of CRISPR-Cas loci

Associations between CRISPR-Cas systems

Having observed that many genomes encode several Cas clusters and that some subtypes are more represented than others in those genomes (Supplementary Figure 4), we searched to identify unexpectedly high or low patterns of co-occurrence between CRISPR-Cas systems. Since genomes are linked by a common evolutionary history, we corrected the associations between systems by the phylogeny using BayesTraits [201]. Since phylogenetic inference of all the bacteria kingdom is very hard, we restricted our analysis to Firmicutes and Proteobacteria, the two clades with more genomes (75% of the total), for which we inferred a set of phylogenetic trees. We used them to test if the co-occurrence of every pair of systems was random (Figure 3.3.A and 3.3.B). In Proteobacteria subtype I-E is negatively associated to both I-B and I-F, while in Firmicutes subtypes II-A and I-B co-occur less than expected. We observed two positive associations in Proteobacteria: subtype I-U with I-C, and IF with III-A/B systems. In Firmicutes, subtype I-B co-occur more than expected with all type III systems. Overall, there were more positive than negative co-occurrences of systems.

According to the patterns of co-occurrence, we encountered clusters that contained proteins from different subtypes. These clusters were often impossible to type, as they included proteins from different subtypes intermingled with multiple copies of homologous components. We hypothesized than these large clusters could represent previously independent systems that physically merged and would therefore indicate potential functional interactions between Cas proteins of differ-

ent CRISPR-Cas systems. Examples of large clusters present in three different bacteria show the physical association of 2 or three subtypes(Figure 3.4.C). We defined large clusters as clusters with more than 12 proteins (Supplementary Figure 5). The protein content of these clusters most often included proteins from type I and type III systems (Figure 3.D). These physical links constitute a new clue to support the hypothesis of interactions between type I and type III systems already suggested by the analysis of the co-occurrence patterns, and one experimental study concerning type I-F and III-B [248].

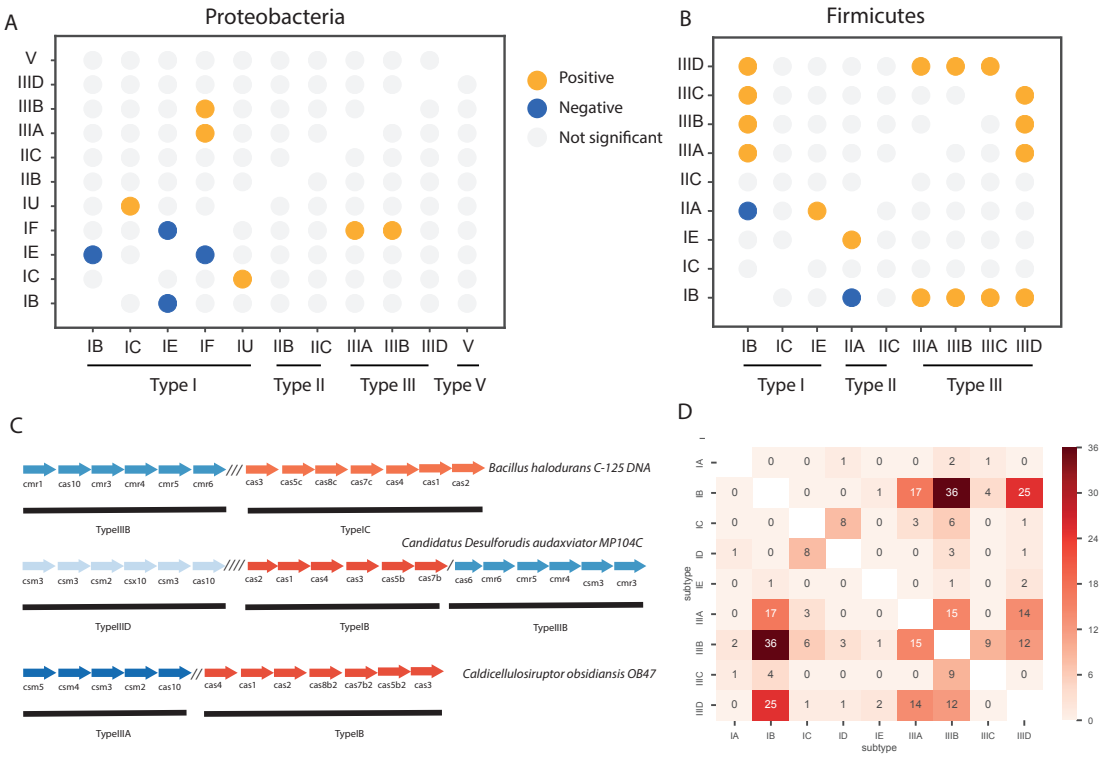


Figure 3.3: The associations between CRISPR-Cas systems.
A) and **B)** Associations between CRISPR-Cas systems in Proteobacteria and Firmicutes. Each circle corresponds to the association between two CRISPR-Cas systems. Associations are represented in grey (not significant), blue (negative), and orange (positive). Only systems present in more than 1% of the genomes in the clade are represented (the others never have significant statistics). **C.** Clusters of Cas proteins detected in three bacteria. Arrows represent genes and are colored by subtypes. Slash (/) represent genes not associated with no hit to cas protein profiles. **D.** Heatmap of proteins subtypes found associated in large clusters (defined as those having more than 12 Cas proteins). Each cell present the number of large clusters where proteins of both subtypes are found.

Characterization of CRISPR arrays according to their association with Cas clusters.

We then investigated if the presence and subtype of Cas clusters influenced the number of repeats of the arrays. There was an association between the size of CRISPR arrays an the presence of Cas genes. Orphan CRISPR arrays are smaller (average 9 repeats) than distant arrays (16), which are smaller than arrays within CRISPR-Cas loci (26, Figure 3.4.A). These trends remained significant when only considering arrays with more than 5 repeats (Tukey HSD, all pairs, $P<0.001$). Plasmidic CRISPR arrays within a locus harbour on average 17 repeats, and are therefore smaller than chromosomal ones (Tukey HSD,$P=0.005$) while plasmidic orphan arrays are not significantly different from chromosomal ones. We then tested if the subtype of a Cas system could influence the size of the CRISPR array. To limit uncertainties of subtype assignment to CRISPR arrays, we only considered CRISPR-Cas loci encoding a single Cas cluster. We observed different numbers of repeats according to subtypes (Supplementary Figure 6). Type IV, V, VI and subtype II-A tend to have short CRISPR arrays (<20 repeats on average). On the other hand, subtype I-A, I-B, I-D have the longest arrays with more than 40 repeats on average. Consequently, the presence, proximity and subtype of Cas clusters impact the number of repeats in CRISPR arrays.

Almost half of the CRISPR arrays are not in CRISPR-Cas loci and many of the remaining are next to complex Cas loci. This means we cannot infer their type based on the contiguity of a single Cas cluster. To further characterize these arrays we built a databank of 3324 unique repeats that we could assign to specific Cas subtypes because they were taken from elements neighboring one and only one Cas cluster (Supplementary Table 2). We performed all pairwise alignments of these repeats and generated an identity score. We then investigated if the subtype of the best hit for a given repeat was a good predictor of the subtype of the repeat. For each repeat of our dataset, we compared the subtype of the best hit and the subtype determined by the Cas of the CRISPR-Cas locus. This method allowed an important number of correct subtype assignments (2689 correct vs 639 incorrect). The accuracy of this simple method is affected by sequence identity: if the best hit has low sequence similarity, then it is a poorer predictor of the Cas type than if sequence similarity is high. We then assessed how the percent identity influenced the quality of the subtype assignment. We performed a logistic regression between the variable for correct subtype assignment (0 or 1) and the identity score. This regression allowed us to set a threshold for the identity score under which we consider that the best hit method is not reliable. Given that in our training dataset, we had an important number of correct subtype assignments, we chose a threshold with a high True Positive Rate (Figure 3.3.C). This threshold corresponds to an identity score of 72%, and in this case, we obtain 2593 correct and 462 incorrect assignments. We used this method to assign subtype to distant and orphan arrays (Figure 3.4.B and Supplementary Figure 4). We could assign

subtypes to 48% of the orphan arrays and 79% of the distant arrays (Figure 3.4.D). Subtypes relative abundance vary between orphan, distant or in loci CRISPR arrays (Supplementary Figure 7) with for example a higher relative frequency of type I-B in distant arrays.

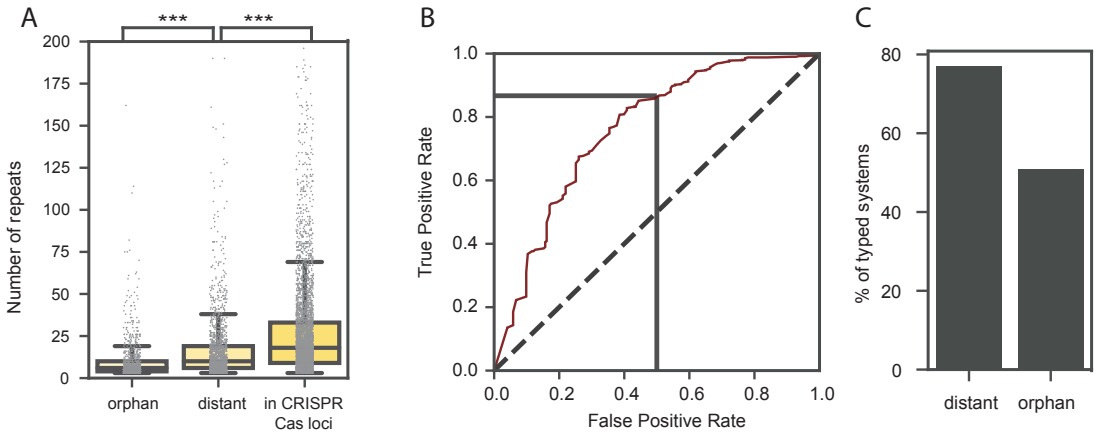


Figure 3.4: Characterization of CRISPR arrays according to their association with Cas clusters.
A. Number of repeats of CRISPR arrays in function of their association (or lack thereof) with Cas clusters (Tukey HSD, all pairs, $P<0.001$). **B.** Predicting CRISPR array subtype using direct repeats. ROC curve (orange) of the logistic regression performed to assess quality of subtype prediction in function of identity score. In grey, the threshold chosen to assign subtype to unknown arrays. **C.** Percentage of typed orphan and distant arrays using the described method.

We then applied this method to test if CRISPR systems encoded on plasmids often match chromosomal Cas. Almost half (48%) of the strains with plasmids harboring CRISPR arrays (but no Cas) are in genomes with a chromosomal Cas cluster (Supplementary Figure 8.A). We assigned a subtype to the plasmidic CRISPR arrays and tested if the array subtype matched the subtype of a chromosomal Cas cluster. We were able to assign a subtype to 37 plasmid CRISPR arrays and that in 18 cases there was a match between the subtype of the plasmid CRISPR array and the chromosomal Cas clusters (Supplementary Figure 8.B). We controlled this result by simulating the expected number of matches between plasmid and chromosomal subtypes. In 1000 simulations, the highest number of matches found was 15 (Supplementary Figure 8.C). While these effectives remain low, this tendency suggests that plasmid CRISPR arrays are more likely to remain in genomes when they are compatible with the chromosomal Cas clusters. A non-exclusive alternative is that these plasmids acquire the CRISPR from their host and spread it in closely related hosts, that are likely to carry similar CRISPR-Cas systems.

3.3 Discussion

We detected CRISPR arrays and Cas clusters at the subtype level in 5563 fully sequenced bacterial genomes. Our detection of Cas clusters underlines an important taxonomic diversity. As our detection takes into account Cas clusters architectures and signature proteins, it provides a robust subtype assignment compared to a previous study where subtypes were only inferred from Cas1 [42]. We found some CRISPR-Cas systems subtypes such as I-B, I-C, II-C, III-B in many clades while others - like subtype II-A or II-B - are clade specific. Such a narrow distribution could be explained by two hypotheses: either such systems co-diverged with their genetic background limiting their functionality in other contexts or they specialized for a defined clade.

Our analysis revealed that 19% of CRISPR arrays encode less than five repeats. These short arrays might be false detections. To limit the impact of potential false positives we controlled our results when necessary using only CRISPR arrays encoding more than five repeats and showed that the observed trends remained the same. These short arrays could also belong to decaying inactive CRISPR-Cas systems, or could result from the dynamics of spacers gain and loss, which in some cases could favour short arrays. The balance of acquisition rate and spacer loss rate is likely influenced by many phenomenon that can vary from species to species, including bacteriophage infection rates, DNA damage and recombination frequencies. Experimental observations on primed adaptation (acquisition of spacers from an MGE already targeted by a spacer in the CRISPR-array) [258] as well as mathematical models predict selective sweeps of highly immune lineages of CRISPR-Cas systems. CRISPR arrays could thus acquire several spacers within a short time-frame. On the other hand, the loss of spacers seems to be a gradual passive phenomenon occurring via homologous recombination between repeats. Short arrays could therefore result from the gradual shrinkage of CRISPR arrays that have not undergone recent acquisition selection events.

While previous studies described the existence of CRISPR-Cas carried by plasmids and phages [235, 177, 99, 236], little was known about the prevalence of such systems. We show that CRISPR-Cas are almost never carried by phages. When CRISPR-Cas are carried by phages, they can nonetheless have an important role as illustrated by how some Vibrio phages use a CRISPR-Cas to escape host innate immunity [236]. CRISPR-Cas systems are more abundant in plasmids, where they remain nevertheless relatively rare. Plasmidic CRISPR arrays are shorter than chromosomal ones, which might indicate that spacer acquisition events might be rarer in this context or that spacer loss is higher (e.g., because higher plasmid copy number increases recombination rates). We observed an intriguing abundance of Type IV systems on plasmids. While it had been previously reported that type IV systems were often found on plasmids [161], we show that they are almost exclusively encoded on them. No experimental studies yet have demonstrated their

activity *in vivo*. As these systems do not encode Cas1 and Cas2, the main proteins for adaptation [161], it is not known how they acquire new spacers. The absence of certain subtypes on plasmids such as type II-A systems also underlines the existence of other mechanisms of horizontal transfer of CRISPR-Cas systems between bacteria.

Our integrated analysis of CRISPR arrays and Cas clusters showed that many systems do not follow the canonical organization of one Cas cluster with one or two CRISPR arrays: 1) approximatively half of CRISPR-Cas loci encode several Cas clusters and CRISPR-arrays. 2) of orphan or distant CRISPR arrays represent 40% of all the CRISPR arrays.

Complex organizations can reflect functional associations between CRISPR-Cas systems of various types. The analysis we performed on subtypes co-occurrences revealed negative and positive associations. We found that type III-B and type I-F are positively associated in Proteobacteria, which could be explained by the ability of type III-B to process and use guide RNAs expressed from a type I-F CRISPR array [247]. We observed many positive association between type I and type III systems. Another form of functional association between type I and type III systems was underlined recently with the discovery of the role of Csm6 [136, 189]. This non-specific RNAse is activated by a small molecule which constitutes an intracellular signal that infection has not been prevented and leads to cell death or dormancy [136, 189, 5]. It was thus hypothesized that these type III systems could constitute a second line of defense when type I CRISPR-Cas systems fail. Finally, type III systems often lack the adaptation module, raising the question of how they acquire new spacers. If they can use the adaptation machinery of type I systems, this would also constitute an explanation for the observed associations.

While complex CRISPR-Cas loci represent one side of the spectrum of CRISPR-Cas organization, the other end e.g. orphan and distant element are also widespread. Distant CRISPR arrays encode fewer spacers than arrays in CRISPR-Cas loci. Such limited size could be explained by a reduced efficiency at incorporating spacers because of a distance to the Cas cluster, which seems unlikely given the molecular mechanisms of immunity which do not require elements adjacency. Therefore, the reasons for this smaller number of repeats mights reside in the origins of the distant arrays. Distant arrays could correspond to a different subtype than the Cas cluster present in the genome, for example as a former active locus where *cas* genes were lost, and would therefore not likely be processed. Distant arrays could correspond to the same subtype but originate from an another locus. As CRISPR arrays in a CRISPR-Cas locus most likely co-evolved with the nearby Cas cluster, they would be less optimized for the distant Cas cluster. Finally, distant arrays could originate from the locus encoding the distant Cas cluster (separated by translocation in the genome for example) and in that case should not be different

from CRISPR arrays in loci.

To further characterize distant and orphan arrays beyond the number of repeats, we provide a new method for assigning a subtype to CRISPR arrays based on their repeats. This simple tool could be incorporated in CRISPR arrays detection programs to provide subtypes when no Cas clusters can be used. It could be especially useful to subtype CRISPR arrays in metagenomics where most of the arrays detected are orphan. Our subtype assignment method is based on repeats of CRISPR arrays within CRISPR-Cas loci. It is therefore probable that we do not sample exhaustively the diversity of CRISPR repeats. The difference in the proportion of the orphan and distant repeats that we could type might indicate that an important pool of repeats might be specific to orphan arrays and thus will be impossible to type using our method.

We present an integrated analysis of CRISPR arrays and Cas clusters bringing new knowledge on CRISPR-Cas systems distribution, organization, co-occurrences and transfers. Overall, our analysis emphasizes the diversity and complexity of such immune systems. As many systems are not easily classable in independent subtypes, a too narrow and strict view of subtypes might be deleterious. Better characterizing complex loci might help understand how certain combinations of components may provide specific advantages to certain loci. A better understanding of this complexity will likely be relevant for the development of novel CRISPR-based technologies.

3.4 Material and methods

Data

We analyzed 5563 complete bacterial genomes retrieved from NCBI RefSeq representing 2437 species of Bacteria (http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/), in November 2016. We retrieved 1943 complete phages genomes from NCBI RefSeq in November 2016.

Prophages detection

We detected 9946 putative prophages in the bacterial genomes using PhageFinder v4.6 as in [269] (http://phage-finder.sourceforge.net/).

CRISPR arrays and Cas clusters detection

Cas clusters were detected with MacSyFinder (version 1.0.2, [1] as in [27] using new Cas profiles and XML-models to consider recently described Type III, IV, V and VI CRISPR-Cas systems [161]. The program is available on the galaxy portal of the Institut Pasteur (https://galaxy.pasteur.fr/). All results are reported in Supplementary Table 1.

We detected CRISPR-arrays using the CRISPR Recognition Tool v1.2 (CRT) [32] with the default parameters except for the maximum length of a CRISPR's repeated region (maxRL) which was set to 50. To limit the number of false positives, we calculated the coefficient of variation of the length of the spacers for each CRISPR array. For CRISPR arrays, this coefficient is expected to be low, as spacers are integrated through mechanisms that ensure specific size [124]. As expected, the number of detected arrays drops when the coefficient of variation rises. To define a threshold above which detected arrays would be considered as false (i.e., detected repeats would not be CRISPR), we analysed the coefficient of variation of CRISPR arrays close to cas genes (here 10kb), which are unlikely to be false. The coefficient of variation is rarely larger than 19 and almost never larger than 28. We therefore chose to remove from the dataset the CRISPR arrays with a coefficient of variation superior to 28. This step removed 318 arrays representing only 4.4% of the dataset. All results are reported in Supplementary Table 2.

Linking CRISPR arrays and Cas clusters

In order to link CRISPR arrays and Cas clusters, we calculated the minimal circular distance between an array and a cluster. When this distance was inferior to 20kb and only one Cas cluster was present, we assigned a subtype to the CRISPR array. This produced two datasets, one of each CRISPR associated to Cas clusters and a second of Cas cluster associated to CRISPR array (Supplementary tables 1 and 2).

We built a databank of repeats assigned to specific subtypes to subtype repeats that could not be assigned trivially because the Cas cluster was either distant from the array or absent from the genome. We analyzed the CRISPR arrays from loci close to one and only one cas cluster (<20 kb), and use them to generate a list of 3324 unique repeats (direct and reverse complement sequences). We quantified the sequence similarity between every pair of such repeats using a global alignment with no gap end penalty and equal gap creation and extension penalties (-3) using the module pairwise2 from Biopython (function align.globalxs). For each repeat of the databank, we generated a categorical variable assessing if the subtype of the best hit in the databank was the same as the one of the tested repeat. We then performed a logistic regression between the identity score of the best hit and the categorical variable assigning correct subtype prediction. We used a ROC curve to choose a threshold with a high True Positive Rate. It corresponds to an identity score of 71.87% to predict correctly a subtype of a repeat based on the best hit in our dataset.

We assigned subtypes to orphan arrays based on this method. For each orphan array, we quantified the sequence similarity with all repeats of the databank using a global alignment with no gap end penalty and equal gap creation and extension penalties (-3) using the module pairwise2 from Biopython (function align.globalxs). We took the best hit among those scores. If the identity score was higher than 71.87%, we assigned to the repeat the subtype of the best hit.

Phylogenetic analyses

We identified the set of families of orthologous genes present in more than 90% of the genomes (when larger than 1 Mb) of two phyla: Firmicutes (1189 genomes), and Proteobacteria (2897 genomes). The genomes were obtained from GenBank's RefSeq dataset as indicated above. The orthologs were identified as reciprocal best hits using end-gap free global alignment, between the proteome of a pivot and each of the other strain's proteomes. Escherichia coli K12 MG1655 and Bacillus subtilis str.168 were used as pivot for each clade. Hits with less than 37% similarity in amino acid sequence and more than 20% difference in protein length were discarded. The persistent genome of each clade was defined as the intersection of pairwise lists of orthologs that were present in at least 90% of the genomes representing 411 families for Firmicutes and 341 for Proteobacteria.

We inferred phylogenetic trees for each phyla from the concatenate of the multiple alignments of the persistent genes obtained with MAFFT v.7.205 (with default options) and BMGE v1.12 (with default options). Missing genes were replaced by stretches of "-" in each multiple alignment. Adding a few - has little impact on phylogeny reconstruction [80]. The trees of the phyla were computed with FastTree version 2.1 using the LG model [218], which had lower AIC than the alternative WAG model in both cases. We made 100 bootstrap trees using phylip's SEQBOOT

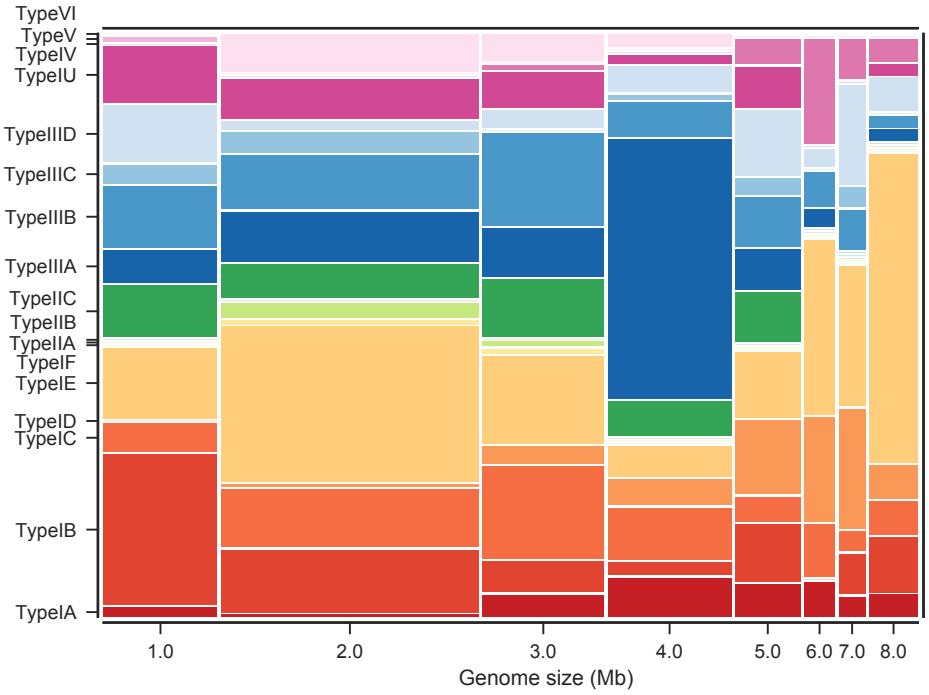
to generate resampled alignments and then using them with FastTree (options n intree1).

We applied BayesTraits v.3. to test the correlations among pairs of traits that adopt a finite number of discrete states, i.e., the presence or absence of peculiar CRISPR-Cas systems [201]. We ran likelihood estimation of two models on 100 trees: independent or dependent evolution of two traits. We performed a likelihood-ratio test on the two models for each of the 100 trees. We validated an association if the median of the p-values of the 100 Likelihood-ratio test was inferior to 0.01 and the p-value of the Fisher exact test on the contingency table of the two traits was inferior to 0.05.

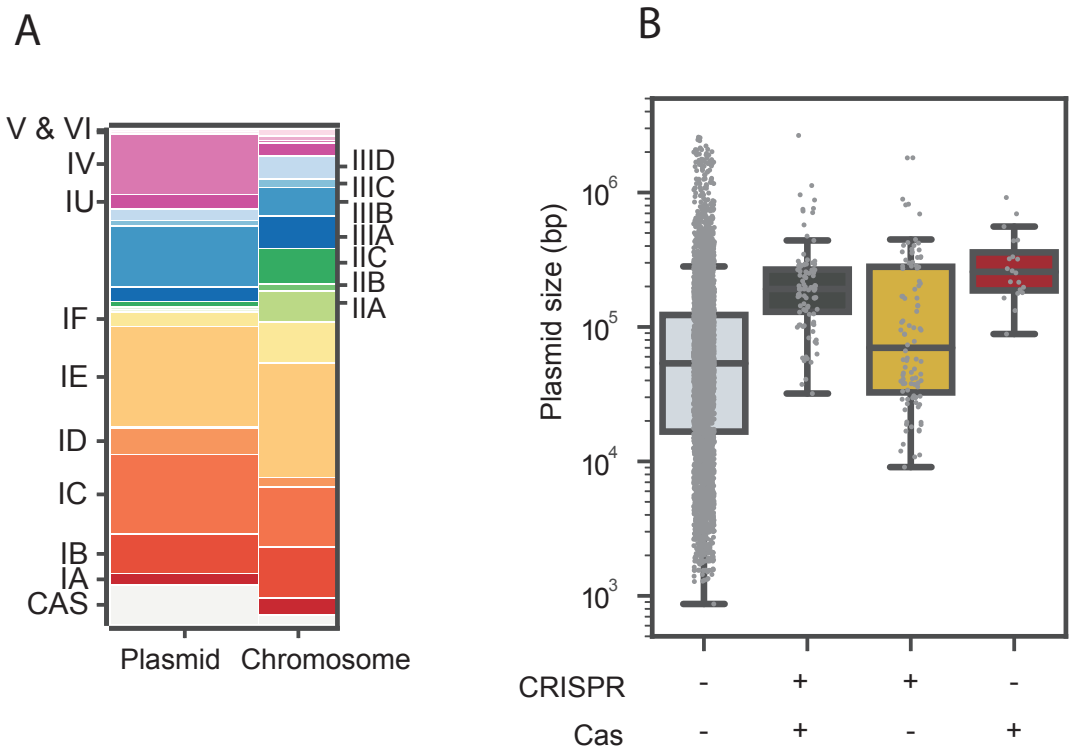
Supplementary materials are available at :
https://gitlab.pasteur.fr/abernhei/Supplementary_Materials_PhD_aude.git

Supplementary materials

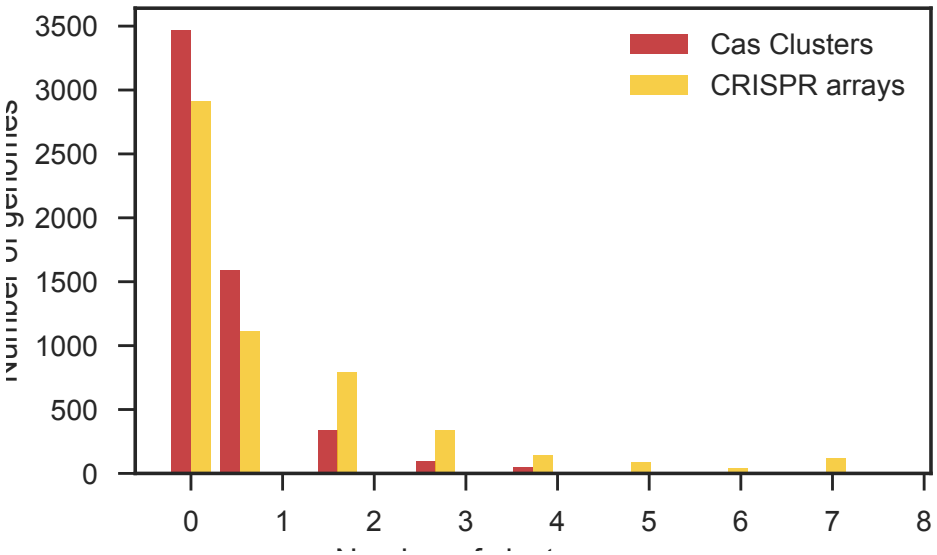
Supplementary Table 1: Cas operons with associated CRISPR arrays
Supplementary Table 2: CRISPR arrays with associated Cas operons



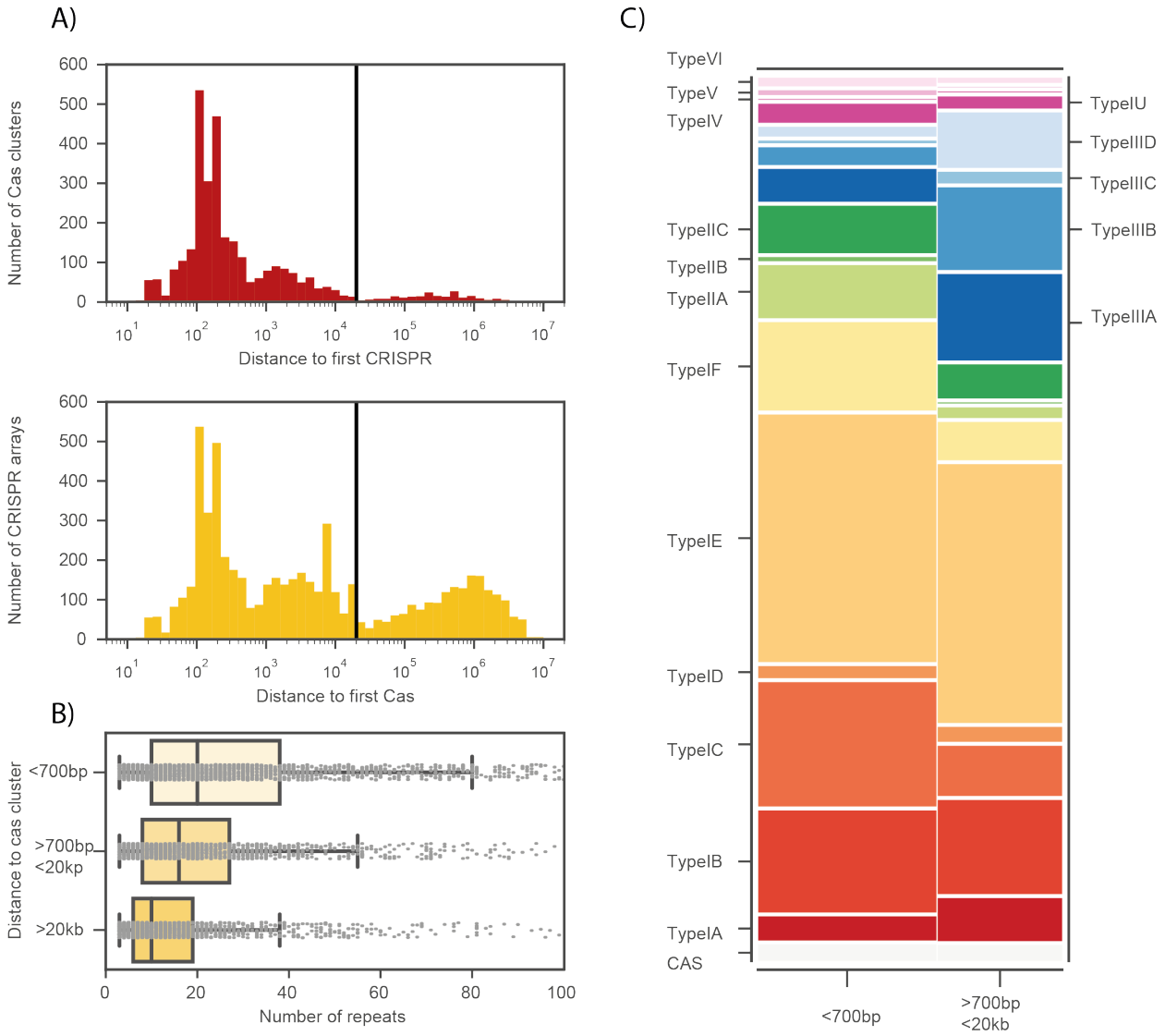
Supplementary Figure 1 : Frequency of subtypes in function of genome size
Each color represent a subtype. X axis represents the distance in Mb.



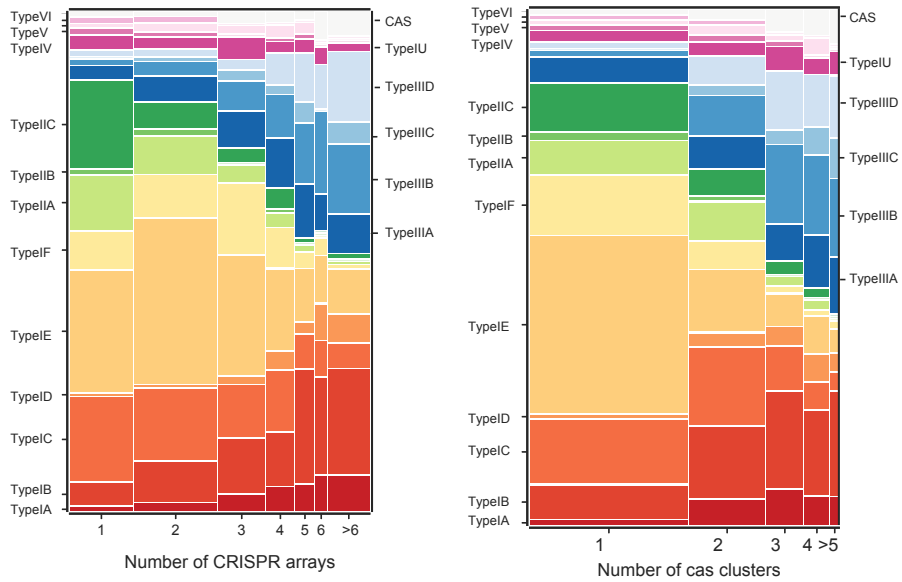
Supplementary Figure 2 : Plasmids and CRISPR-Cas systems
A. Plasmidic Cas subtypes. Frequency of chromosomal and plasmidic Cas subtypes (Chi 2 on the independence of variables in a contingency table, P<0.001). CAS means that no specific subtype could be assigned. **B.** Plasmid size in function of the presence of CRISPR-Cas systems (Tukey Kramer CRISPR, P=0.033, Cas, P=0.002, CRISPR-Cas, P<0.0001).



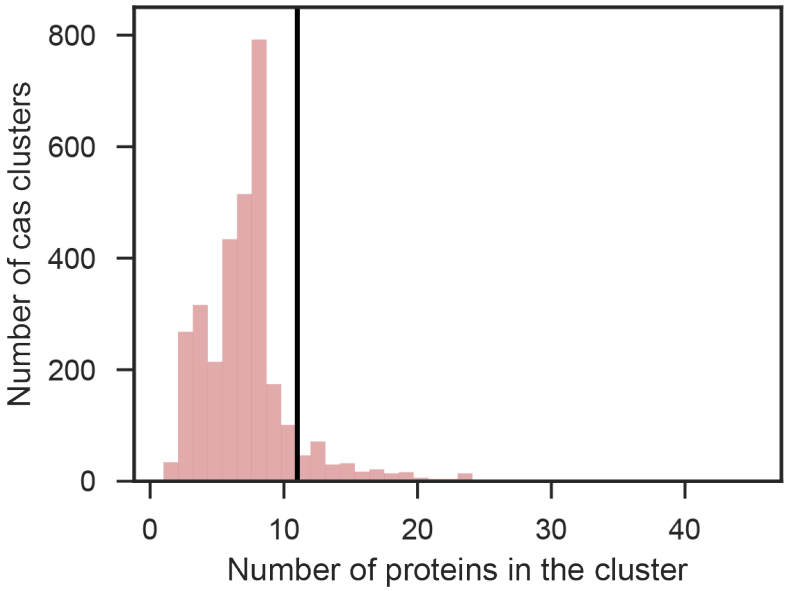
Supplementary Figure 3 : Number of CRISPR-arrays and cas clusters in bacterial genomes.



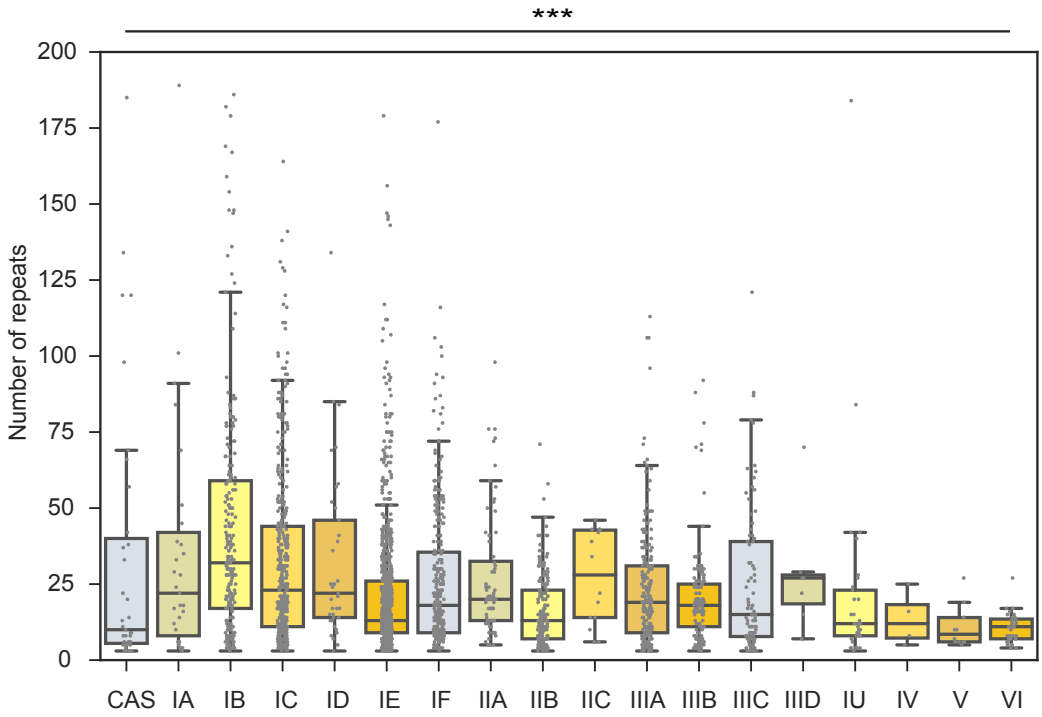
Supplementary Figure 3: Comparison of three groups of CRISPR arrays and Cas clusters
A. Distance to the first CRISPR array or first Cas cluster. Vertical line corresponds to the distance cut-off (20kb) **B**. Number of repeats in different groups of CRISPR-arrays. **C.** Subtypes associated to the closest Cas clusters in the different groups.



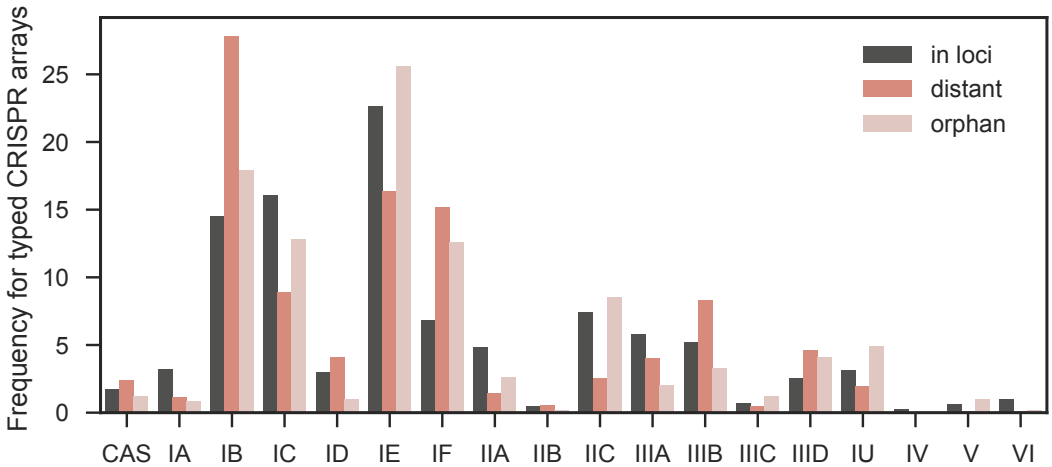
Supplementary Figure 4 : Subtypes in function of number of Cas clusters and CRISPR arrays



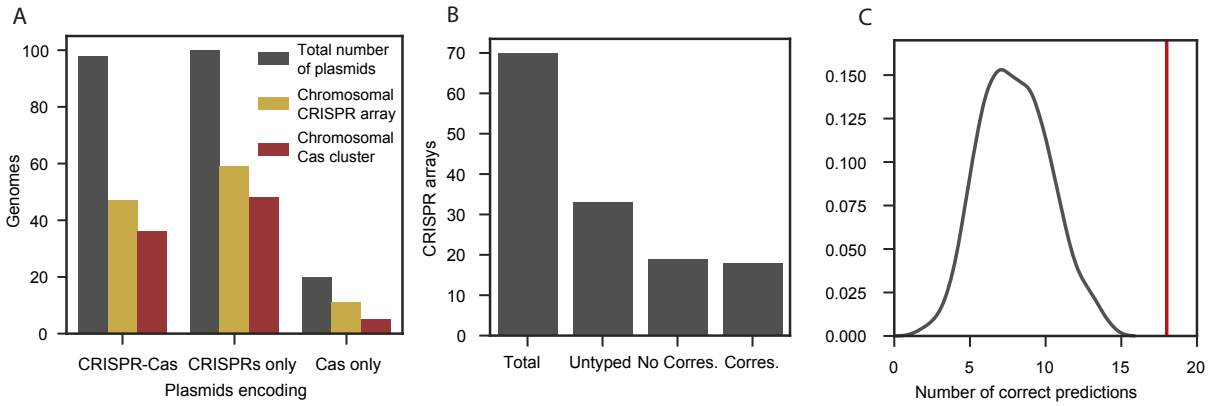
Supplementary Figure 5 : Number of proteins in cas clusters
The vertical line corresponds to the cut-off (12 proteins) to define “large clusters”



Supplementary Figure 6 : Number of repeats of CRISPR arrays per subtype
Number of repeats of CRISPR arrays in function of the subtype of the closest Cas clusters. Only CRISPR arrays associated with one specific subtype and within 20kb of a Cas cluster were analysed (ANOVA, P<0.001).



Supplementary Figure 7 : Subtypes abundance in CRISPR arrays
CRISPR subtype was determined by the subtype of the associated Cas cluster for the “in loci”. Distant and orphan CRISPR arrays were subtyped using the method described in the article, only based on their repeats. The frequencies of the types CRISPR arrays are represented which correspond to 100% of the in loci arrays, 77% of the distant and 51% of the orphan subtypes (Chi 2 on the independence of variables in a contingency table, P<0.001).



Supplementary Figure 8 : Subtype associations between plasmidic CRISPR arrays and associated chromosomal Cas clusters.

A. CRISPR-Cas content of genomes with plasmids encoding CRISPR arrays and/or Cas clusters. **B.** Subtype correspondence between plasmidic CRISPR array and chromosomal Cas clusters. Subtype was assigned to CRISPR arrays encoded on plasmids. Correspondence was defined as such: If the assigned subtype of the array matched one of the Cas clusters present in the chromosome **C**. Control for plasmidic CRISPR array subtype correspondence to Cas cluster chromosomal one. Vector corresponding to the CRISPR subtypes was randomized and number of correspondence with the associated vector of Cas clusters was computed. The distribution of the number of correct predictions for 1000 randomizations is presented. The number of observed correspondence in our dataset is represented by the vertical red line.

Box 2: Major points of Chapter 3

- **First quantitative integrated analysis** of CRISPR arrays and Cas clusters.
- CRISPR-Cas subtypes are **distributed unevenly** among bacterial genomes. Some are clade specific while others are widespread. CRISPR-Cas systems are **almost never carried by phages** but many are **encoded on plasmids**.
- **Orphan and distant arrays represent 40% of detected CRISPR arrays**. They harbour less repeats that CRISPR arrays adjacent to Cas clusters.
- Many CRISPR-Cas loci present **complex architectures** with several CRISPR arrays and/or Cas clusters.
- **Type I and Type III CRISPR-Cas systems often co-occur** in bacterial genomes underlying potential synergistic interactions.
- New **method to assign a subtype to CRISPR arrays ~~only~~ using the sequence of the repeat**. This method was applied to show that plasmids tend to encode CRISPR arrays compatible with the Cas systems encoded by their host genomes. It could be particularly useful in metagenomics.

Chapter 4

Interactions between DNA repair and CRISPR-Cas systems as a cause for the sparse distribution of these systems in bacteria

The work presented in this chapter is still ongoing. Some of the results are preliminary and might be subject to changes.

Aude Bernheim[1,2,3,4], David Bikard[3], Marie Touchon[1,2], Eduardo P.C. Rocha[1,2,*]

[1]Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue Dr Roux, Paris, 75015, France
[2]CNRS, UMR3525, 25-28 rue Dr. Roux, Paris, 75015, France
[3]Synthetic Biology Group, Institut Pasteur, 25-28 rue Dr. Roux, Paris, 75015, France
[4]AgroParisTech, F-75005 Paris, France
[*] Provisional order of the authors

Abstract

The distribution of CRISPR-Cas systems is sparse in many bacterial phyla. Considering their important role as an adaptive immune system, the absence of these systems in more than half of bacteria is puzzling. Here, we investigate the possibility that the success of CRISPR-Cas acquisition by horizontal gene transfer is partly determined by the interactions of these systems with the genetic background of the host. More specifically, we analyze the co-occurrence patterns between CRISPR-Cas systems and the DNA repair systems of bacteria, especially in terms of the functions involved in homologous recombination and handling double strand breaks, like non-homologous end joining. We identify the few previously

studied cases of positive and negative interactions between the systems and show that there are many other positive and negative patterns of co-occurrence between DNA repair and CRISPR-Cas systems. This study provides a novel explanation for the absence of CRISPR-Cas systems in the majority of bacteria and opens numerous avenues for further experimental research on interactions between these systems and bacterial DNA repair pathways.

4.1 Introduction

CRISPR-Cas systems are an adaptive immune systems of bacteria and archaea. They are present in less than 50% of bacteria and some phyla are completely devoid of them [161, 42]. Given their role as an immune system and the proofs of their frequent horizontal transfer [90, 47], the absence of systems in large clades enduring horizontal gene transfer and phage predation remains intriguing. As a point of comparison, there are on average two restriction-modifications systems per genome [197]. Several hypotheses have been put forward to explain this relative scarcity of CRISPR-Cas systems. First, autoimmunity: the acquisition of a self-targeting spacer leads in the vast majority of cases to cell death [110]. Second, harboring general defenses like restriction modification or surface modification can be more advantageous than encoding a specialized defense system like CRISPR-Cas [296]. Third, by limiting horizontal gene transfer (HGT), CRISPR-Cas systems can prevent the uptake of advantageous mobile genetic elements [128]. None of these explanations if fully satisfactory. First, there is not clear reason why the cost of auto-immunity should vary between clades. Second, it isn't clear why Archaea would always select for specialized and general defense systems and so many Bacteria would only select for the latter. Finally, the costs of preventing HGT is general for all defense systems.

Here, we propose a novel and complementary explanation: incompatibility with the bacterium genetic background. CRISPR-Cas systems are constantly lost and gained again through HGT [128]. We hypothesize that the success of the transfer will depend on the genetic background. Indeed, from a mechanistic point of view, some systems require host factors to be fully functional, like the dependence of the type I-E system on the IHF (Integration Host Factor) to integrate new spacers [192, 301]. In this case, one would expect to find frequent co-occurrence between the CRISPR-Cas system and the host factors (whenever the latter are not ubiquitous). Incompatibilities between CRISPR-Cas systems and other encoded pathways could also prevent the fixation of transferred CRISPR-Cas systems. If a given trait is affected by the presence of a given type of CRISPR-Cas system, or renders the system inefficient, then the two should rarely co-occur.

To test the hypothesis of the importance of the genetic background in the distribution of CRISPR-Cas systems, we focused on DNA repair. Our choice was motivated by several reasons. First, CRISPR-Cas systems and DNA repair pathways share the same substrate, DNA, which may result on competition for the same substrate. Second, by potentially being able to repair breaks generated by CRISPR-Cas systems, DNA repair pathways could limit CRISPR-Cas efficiency. Third, some DNA repair proteins have been shown to interact with CRISPR-Cas systems and even play a role in CRISPR-Cas adaptation leading to potential synergistic interactions [149, 122, 110]. For example, in type I-E systems, RecBCD

DNA degradation generates small pieces of DNA that are used as prespacers by the adaptation machinery [149].

We focused our analysis on DNA repair pathways and proteins which we hypothesized could interact with CRISPR-Cas systems. DNA repair pathways could either impact interference of CRISPR-Cas systems by repairing double strand breaks (DSBs) or play a role in adaptation by for example producing prespacers in a manner similar as the one described for RecBCD and type I-E systems. Therefore, we investigated associations with proteins involved in homologous recombination (HR) and Non-Homologous End Joining (NHEJ) pathway which can repair DSB without any template [38]. In HR, the main DSB repair complexes AddAB, RecBCD, AdnAB [175, 15, 251] which exhibit helicase and nuclease activities, process DSB and then recruit RecA [299]. Other pathways have been described to repair DSB in specific genetic backgrounds. They involve: RecOR and RecF, which bind to DNA and recruit RecA; the helicase RecQ, the nuclease RecJ and RecN which tethers DNA [175, 188]. The resolvases RuvAB, RuvC, RecU and RecG carry out the post-synaptic phase through Holliday junctions resolution [175]. We also examined regulators of homologous recombination: SbcB, SbcEF, SbcCD, PnPase and RecX [54]. Finally, we also considered SOS response through LexA, which repression on many genes is alleviated during the SOS response.

In this study, we examined the patterns of co-occurrence of DNA repair pathways and CRISPR-Cas systems in bacterial genomes to test if they were independently distributed. If one subtype of CRISPR-Cas systems interacts synergistically with a specific DNA repair pathway, then both systems should co-occur more often in bacterial genomes that what is expected under independent distributions. Conversely, antagonistic systems should co-occur less.

4.2 Results and Discussion

Distribution of DNA repair pathways in bacterial genomes

We detected CRISPR-Cas systems and proteins involved in DNA repair in 5563 fully sequenced bacterial genomes (Supplementary Table 1). The distribution of DNA repair pathways and proteins detected confirmed previous analysis [73, 227, 56]. Several systems or proteins are nearly ubiquitous. They include RecA, the resolvases RuvAB and RecG, and the pre-synaptic system RecOR. We could detect them in more than 96% of the genomes of the dataset (Figure1). Genomes of species lacking RecA were on average half the size of the others and were much more likely to lack the other systems too. These systems represent the nearly ubiquitous toolkit of homologous recombination in bacteria.

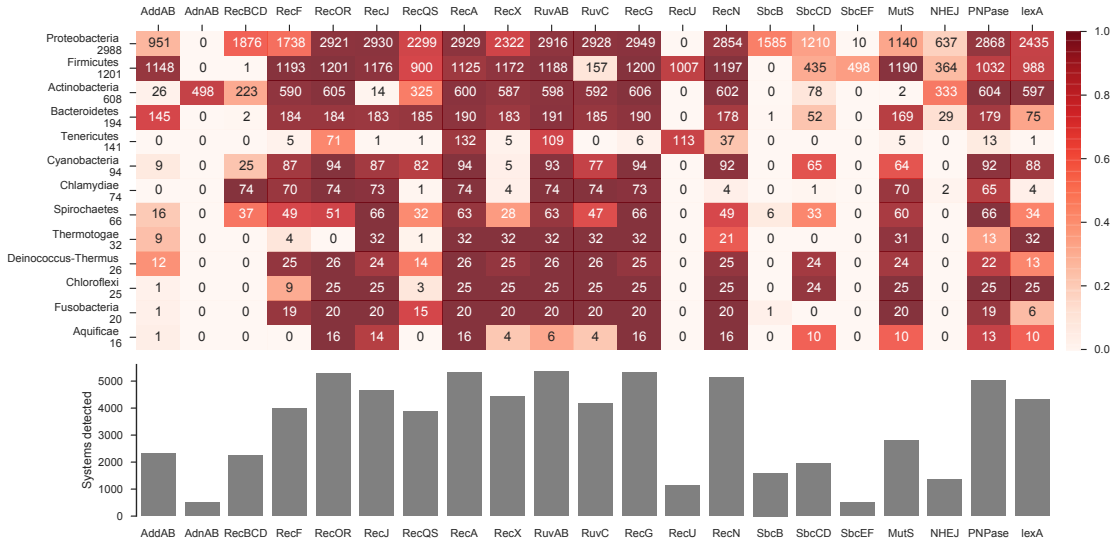


Figure 4.1: Distribution of DNA repair pathways in bacterial genomes. The top panel represents the taxonomical distribution of different DNA repair sets of proteins indicated on the x-axis. Clades are ordered by number of genomes present in the dataset which are indicated on the y axis. Each cell represents the number of systems detected for the clade and is colored proportionally to the frequency of the system in the clade, the darker, the more frequent. The bottom panel is the total number of systems detected in the dataset.

Certain systems are not ubiquitous individually, because there are several different epistatic groups of proteins with similar functions (Figure 1). For example, genomes encoded either RecBCD or AddAB or AdnAB. RecBCD is virtually absent from certain clades like the Firmicutes or the Bacteriodetes, and AdnAB is restricted to Actinobacteria. One should note that some actinobacteria have both AdnAB and RecBCD, but the latter seems to be involved in single-stranded annealing and not in homologous recombination [100]. RecU and RuvC, which are

both resolvases, also have complementary patterns of occurrence (the former is only found in Fimicutes and Tenericutes). It is interesting to notice that some clades like the Tenericutes and Thermotogae harbor few identifiable DNA repair proteins which could be linked to their small genome size. NHEJ is present in 25% of bacterial genomes. It is more abundant in Actinobacteria where 55% of the genomes encode NHEJ.

Interactions between CRISPR-Cas systems and DNA repair in Proteobacteria and Firmicutes

We analysed the patterns of co-occurrence of DNA repair pathways and CRISPR-Cas systems to test whether they were distributed independently. When studying co-occurrences of genes, it is important to consider that genomes are linked by a common evolutionary history and that statistics should be corrected by the phylogeny [79, 200, 201]. As the presence of CRISPR-Cas systems is variable within species, we decided to analyse the patterns of co-occurrences at the genome level (i.e., using multiple genomes for the same species when available). This increases the dataset, at the cost of increasing the phylogenetic association between taxa. To control for this last effect, we built phylogenetic trees associating all genomes. It quickly appeared that reconstructing a trustworthy phylogeny on our whole dataset of 5563 genomes would not be possible. We thus decided to focus our analysis on the two best-studied clades: Firmicutes and Proteobacteria. These were the only phyla with genomes encoding enough CRISPR-Cas systems from various types to perform robust statistical analyses (respectively 293 and 929 type I, 218 and 119 type II, 109 and 62 type III). We built the trees for Firmicutes and Proteobacteria, separately, and used them to test the co-occurrence of systems with BayesTraits [201]. The significant associations found in this analysis are shown in Figure 2.

First, we detected a positive association in Proteobacteria between RecBCD and type I-E CRISPR-Cas systems. This association is consistent with the synergistic interaction between both pathways that was previously identified and studied in *E. coli* [149]. Naïve adaptation in type I-E systems relies on RecBCD to provide prespacers for the adaptation machinery. We also observed a negative association between type II-A CRISPR-Cas systems and NHEJ in Firmicutes. We reported this negative association in a previous study in which we provided experimental evidence showing that type II-A CRISPR-Cas systems likely inhibit NHEJ repair [27]. Another interaction between DNA repair proteins and CRISPR-Cas systems reported recently is the contribution of RecG to primed adaptation (acquisition of news spacers from an MGE already targeted by a spacer present in the CRISPR array) in type I-E and I-F systems [122, 110]. However, we were unable to detect this interaction in our analysis. This could be explained by the ubiquituous presence of RecG (>96% of the genomes in the dataset). Hence, in this case our analysis suggests that there is no reciprocal co-occurrence of type I-E and I-F

with RecG, just a functional dependence of the former on the latter. Beyond the handful of interactions previously reported, we observed 105 new significant associations (Figure 2). Some genes lack any significant co-occurrence and they were removed from the figures. This is the case of most almost ubiquitous components of the HR machinery, including the above mentioned RecG, but also, RuvAB, and RecA.

We first focused on the global patterns of these associations. The first striking observation is that the co-occurrence patterns are not the same in Firmicutes and Proteobacteria. An important factor to explain these results is that CRISPR-Cas systems are not equally distributed in those clades, with for example only 39 type I-E/F systems in Firmicutes and 773 in Proteobacteria. Similarly, DNA repair pathways are also distributed unevenly among phyla (Figure 1). Evolutionary associations can only be detected when a sufficient number of systems are present in the clade under study. This likely explains some of the discrepancies between Firmicutes and Proteobacteria (Figure 2.A). However, for some DNA repair proteins like AddAB, the patterns are almost opposite in the two phyla.

It is important to highlight that in many cases, these evolutionary associations likely do not indicate direct mechanistic interactions but could result from indirect associations with other traits for example genome size. Other potential indirect effects could from the preferred associations of some DNA repair pathways. Some frequently co-occur with one another like RuvAB and RuvC or on the contrary present a complementary distribution (bacteria have one or the other) like AddAB and RecBCD. A direct consequence of the associations between DNA repair pathways is that the interaction of a CRISPR-Cas system with a specific DNA repair pathway will impact co-occurrence patterns with multiple DNA repair pathways.

For instance, RecU and RuvC exhibit a complementary distribution in Firmicutes and as expected present the exact opposite association pattern with CRISPR-Cas systems. RuvC is positively associated with Type I and III systems, but negatively associated with type II-A systems while RecU presents the opposite pattern. Similarly, Proteobacteria encode either the AddAB or the RecBCD presynaptic pathway. These systems present opposite associations with CRISPR-Cas systems in this phylum. Conversely, *recBCD*, *sbcCD*, *sbcB* and *lexA* tend to be encoded in the same genomes and thus show similar patterns of co-occurrence with CRISPR-Cas systems.

To facilitate the analysis of these evolutionary associations we performed a hierarchical clustering of the CRISPR-Cas systems based on their interactions with DNA repair pathways (Figure 3). This analysis revealed some expected clusters like the one including type I-E and I-F systems in Proteobacteria (Figure 3.A). Indeed, those two systems are very similar in terms of molecular mechanisms.

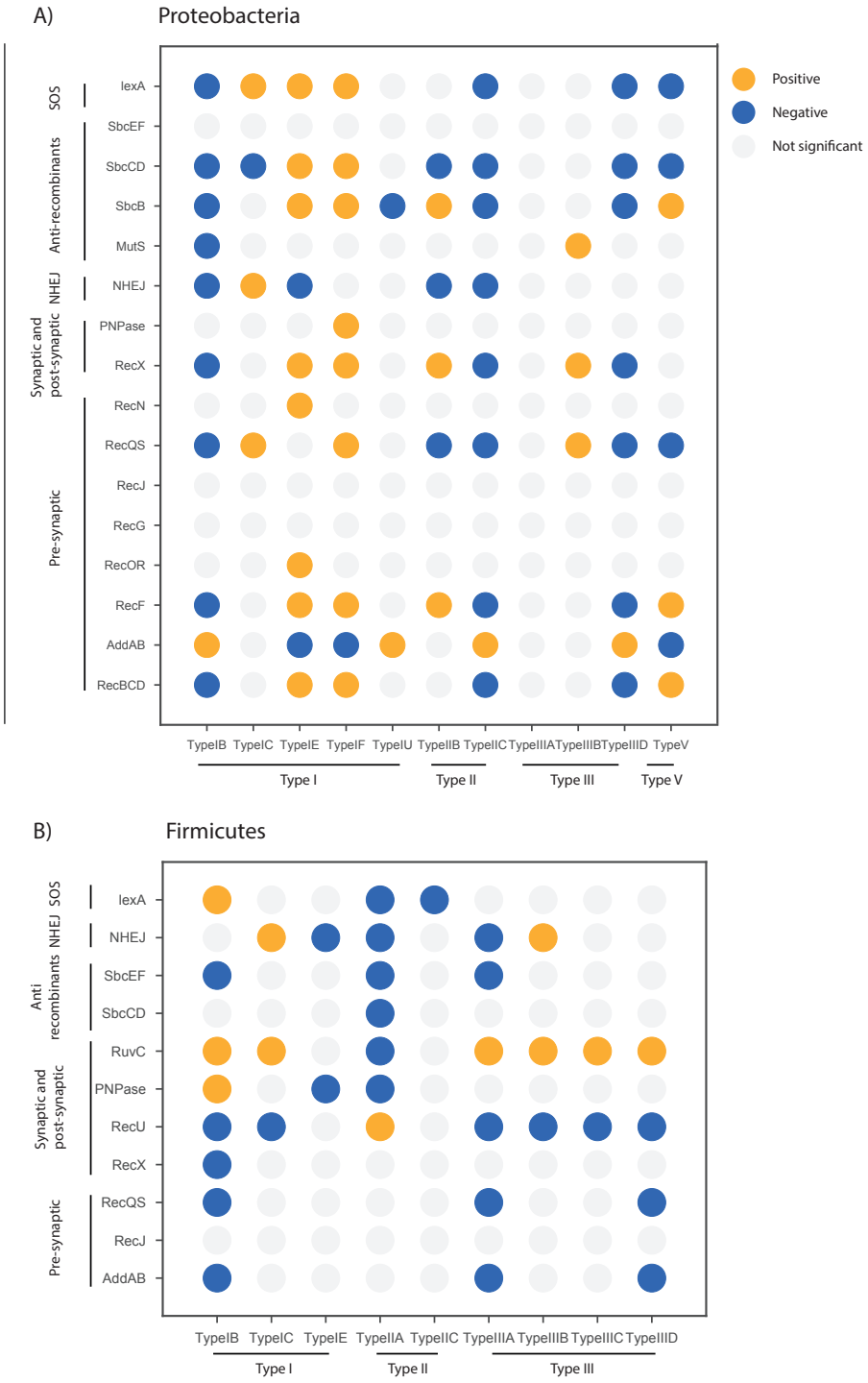


Figure 4.2: Associations between CRISPR-Cas systems and DNA repair in Proteobacteria and Firmicutes.
Each circle corresponds to the association between a CRISPR-Cas system on the x-axis and a DNA repair pathway in the y axis. Grey represent no significant association, blue negative association and orange positive one (Fisher Exact Test $P<0.05$, median of a 100 likelihood ratio tests <0.01). Only systems present in more than 1% and in less than 99% of the total number of genomes in the clade are represented. The top panel depict associations in Proteobacteria and the bottom one in Firmicutes

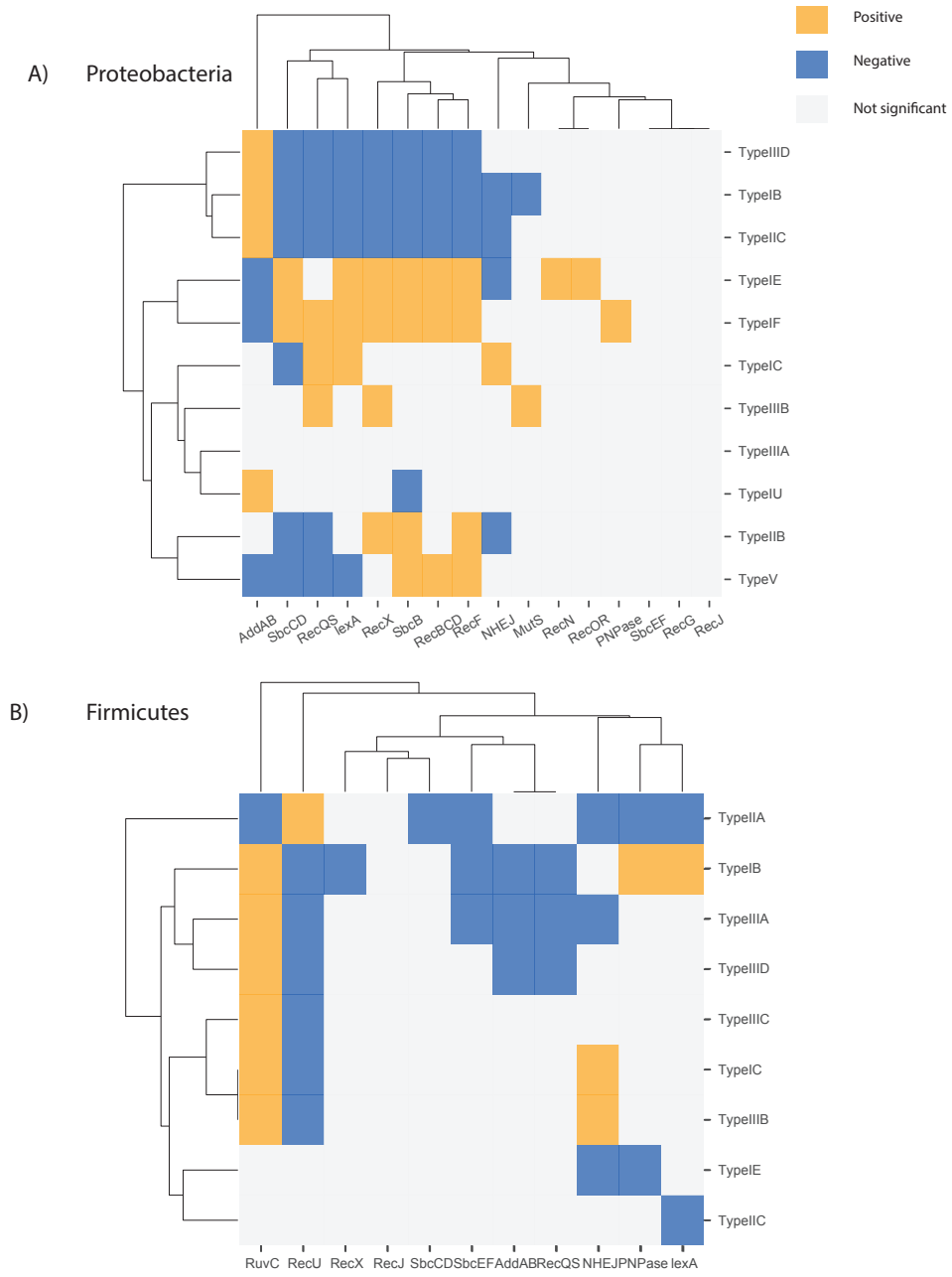


Figure 4.3: Hierarchical clustering of CRISPR-Cas systems by their associations with DNA repair pathways.
Each square corresponds to the association between a CRISPR-Cas system on the y-axis and a DNA repair pathway in the x axis. Grey represents no significant association, blue represents a negative association and orange a positive one (Fisher Exact Test $P<0.05$, median of a 100 likelihood ratio tests <0.01). Only systems present in more than 1% and in less than 99% of the total number of genomes in the clade are represented. The top panel depict associations in Proteobacteria and the bottom one in Firmicutes.

However, some clusters were more surprising, like the ones grouping together type I-B, II-C and III-D in Proteobacteria. Those subtypes belong to different types of CRISPR-Cas systems and differ a lot in Cas proteins content, leaving no specific reasons for such common patterns. Also surprising, some systems very similar at the molecular level show very different patterns like type II-A and type II-C in Firmicutes (Figure 3.B). The hierarchical clustering thus underlines the diversity and subtype specificity of CRISPR-Cas systems and the correlations between DNA repair systems.

The specific interactions will have to be studied in details in the future. As setting up experimental models to validate these interactions would take a lot of time, understanding the underlying molecular details will have to be a community work. However, this analysis provides a roadmap for which interactions to focus on.

Interactions between CRISPR-Cas systems and DNA repair pathways shape CRISPR-Cas systems distribution in bacterial genomes

Given the number of significant associations between CRISPR-Cas and DNA repair systems, we propose a scenario for the impact of these associations on the distribution of CRISPR-Cas systems (Figure 4). CRISPR-Cas systems are subject to frequent horizontal gene transfer. When they are introduced in a novel bacterium, their effect on fitness will depend on a number of factors, such as the impact of phage predation on the population, the expression of *cas* genes, or the presence of other defense systems. And it will also depend on the genetic background, especially in what concerns functions that are associated with DNA repair. Diverse bacteria encode different DNA repair pathway and these may have diverse degrees of compatibility to specific CRISPR-Cas systems. On one extreme, the CRISPR-Cas system may depend on the existence of a DNA repair pathway to be fully functional. This seems to be the case of the type I-E and RecBCD [149]. On the other extreme, there may be strong incompatibility between the system and a DNA repair pathway. This seems to be the case for type II-A and NHEJ, because the former affects the function of the latter [27]. Upon a transfer of a new CRISPR-Cas system, the interactions with the extant DNA repair pathways will significantly contribute to the overall change in fitness resulting from the acquisition of the system, thus driving its loss or fixation in the lineage. A consequence of this process, is that bacteria with different DNA repair pathways will end up encoding different CRISPR-Cas systems. Hence, our results may contribute to explain the scattered distribution of these immune systems in bacteria.

Beyond informing what shapes the distribution of CRISPR-Cas systems in bacterial genomes, the study of DNA repair pathways interactions with CRISPR-Cas systems could lead to new findings on CRISPR biology. Given the essential role of RecBCD in adaptation for type I-E systems and more generally on how DSB constitute a preferred source for prespacer production, the reported associations pave

the way for experimental studies to better understand adaptation mechanisms in different subtypes of CRISPR-Cas systems. Moreover, the interactions between CRISPR-Cas systems and DNA repair pathways are at the heart of CRISPR based technologies both for genome editing and sequence specific antimicrobials [57]. While for genome editing, efficient repair is sought, CRISPR based antimicrobials rely on the impossibility for bacteria to repair efficiently their genome. A better understanding of the natural interactions in bacteria could therefore help improve such microbial technologies by allowing for example a fine tuning in the choice of subtype to use given a specific genetic context.

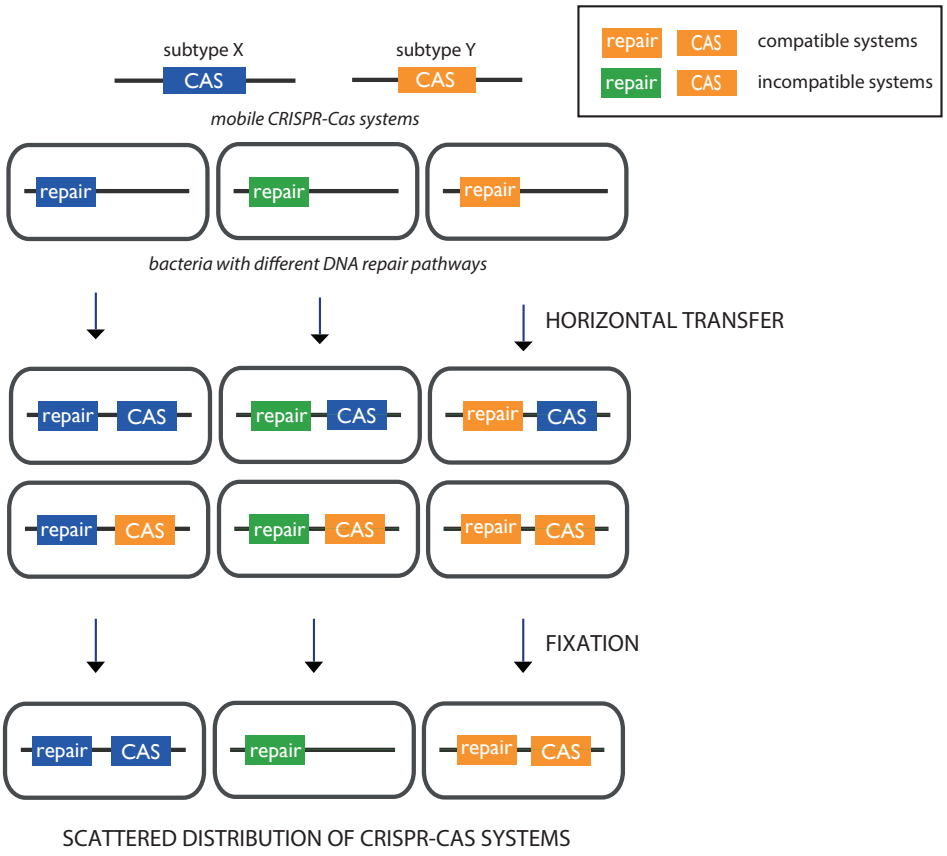


Figure 4.4: Consequences of the interactions between DNA repair pathways and CRISPR-Cas systems on CRISPR-Cas system distribution in bacterial genomes.
Different subtypes of CRISPR-Cas systems have different compatibilities with specific DNA repair pathways. When a CRISPR-Cas system is integrated in a bacterial genome, it will be fixed or lost based on this compatibility.

4.3 Material and methods

Data

We analyzed 5563 complete genomes retrieved from NCBI RefSeq (`ftp://ftp.ncbi.nih.gov/genomes/`), last accessed in November 2016) representing 2437 species of Bacteria.

Detection of CRISPR-Cas systems

CRISPR-Cas systems were detected with MacSyFinder (version 1.0.2, [1]) as in [27] using Cas profiles and system XML-models (Cas Finder available on the galaxy portal of Institut Pasteur) accounting for the recently described Type III, IV, V and VI CRISPR-Cas systems [161]. All results are reported in Supplementary Table 1.

Construction of protein profiles

We build protein profiles for several proteins that either lacked profiles in the public database or had profiles which were not specific enough. The procedure was the same for each protein. First, we collected a set of protein sequences. The homologous proteins were then aligned using MAAFT (default parameters, mode auto). Multiple alignments were then manually curated using Seaview v 4.6.2. Finally, the multiple alignments were used to produce protein profiles with hmmbuild from the HMMer suite version 3.1. For AddA and AddB, we first obtained a list of representative proteins from different clades as described in [56](list in Supplementary Table 3). As known functional homologs in Epsilonproteobacteria were not detected by this customed profiles, two specific profiles to detect AddA and AddB in Epsilonproteobacteria were built using sequences from [6]. For AdnA, AdnB, SbcB, SbcE, curated proteins from Uniprot were used as a starter for a Blast (Blast-p NCBI, may 2016) against the non redundant protein sequences database of NCBI . All hits belonging to different clades among the 250 best hits with more than 40% identity were selected and aligned as described above.

Detection of DNA repair pathways

We used MacSyFinder to detect DNA repair pathways and DNA repair proteins [1].We retrieved protein profiles from TIGRFAM or built custom profiles when no profiles existed (AdnA, AdnB, SbcB, SbcE) or when detection using TIGRFAM profiles missed known homologs (AddA, AddB). We built MacSyFinder models for these systems (Supplementary Text 1). We compared these results to a previous analysis using other methods in smaller sets of genomes [227, 56].

Phylogenetic analyses

We built persistent genomes that is to say all families of orthologous genes that were present in more than 90% of the genomes, for 1189 Firmicutes and 2897 Proteobacteria genomes larger than 1 Mb available in the GenBank RefSeq dataset indicated above. A list of orthologs was identified as reciprocal best hits using end-gap free global alignment, between the proteome of a pivot and each of the other strain's proteomes. Escherichia coli K12 MG1655 and Bacillus subtilis str.168 were used as pivot for each clade. Hits with less than 37% similarity in amino acid sequence and more than 20% difference in protein length were discarded. The persistent genome of each clade was defined as the intersection of pairwise lists of orthologs that were present in at least 90% of the genomes representing 411 families for Firmicutes and 341 for Proteobacteria. We made phylogenetic trees for each clade from the concatenate of the multiple alignments of the persistent genes obtained with MAFFT v.7.205 (with default options) and BMGE v1.12 (with default options). Missing genes have been replaced by stretches of "-" in each multiple alignment. Adding - has little impact phylogeny reconstruction as long as these are not very numerous [80]. Each clade tree was computed with FastTree version 2.1 under LG model [218]. In both cases, LG model minimized the AIC compared to the WAG model. We made 100 bootstraps to assess the robustness of the phylogenetic reconstruction using phylip's SEQBOOT to generate resampled alignments and the n intree1 options of FastTree.

We applied BayesTraits v.3.0 [201] to test the correlations among pairs of categorical traits. In our case, these traits were all binary (presence or absence of peculiar CRISPR-Cas and DNA repair syste). We estimated the likelihood of the presence or absence of the two traits using two models: one where it is hypothesized that the discrete traits evolved independently and one where the characters would have evolved in a correlated manner. We performed likelihood-ratio tests on the 100 trees provided by the bootstraps to account for phylogenetic uncertainty. We considered that an association was valid if the median of those 100 likelihood ratio tests was inferior to 0.01 and if the p-value of the Fisher Exact test was inferior to 0.05.

Clustering

Each association was assigned 0,-1 or 1 representing not significant, negative and positive. The matrix of these associations was clustered using hierarchical clustering (clustermap function from the seaborn package in Python 2.7 with default parameters). The function uses the Nearest Neighbor Algorithm method to form clusters.

Supplementary materials are available at :
https://gitlab.pasteur.fr/abernhei/Supplementary_Materials_PhD_aude.git

Box 3: Major points of Chapter 4

- **Detection of positive and negative associations between CRISPR-Cas systems and DNA repair pathways** by analyzing co-occurence patterns in bacterial genomes.
- **Identification** of the few **previously studied interactions**.
- Associations vary for different subtypes of CRISPR-Cas systems belonging to the same type.
- We propose a **scenario to explain the scattered distribution** of CRISPR-Cas systems. Different subtypes of CRISPR-Cas systems have different compatibilities with specific DNA repair pathways. When a CRISPR-Cas system is transfered to a new host, it will be fixed or lost based on this compatibility.

Chapter 5

Inhibition of NHEJ repair by type II-A CRISPR-Cas systems

This chapter presents the study of a specific interaction : NHEJ and type II-A CRISPR-Cas systems. The manuscript below was accepted for publication in Nature Communications.

Inhibition of NHEJ repair by type II-A CRISPR-Cas systems

Aude Bernheim[1,2,3,4], Alicia Calvo Villamanan[3], Clovis Basier[3], Lun Cui[3], Eduardo PC Rocha[1,2], Marie Touchon[1,2,*], David Bikard[3,*]

[1] Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue Dr Roux, Paris, 75015, France
[2] CNRS, UMR3525, 25-28 rue Dr. Roux, Paris, 75015, France
[3] Synthetic Biology Group, Institut Pasteur, 25-28 rue Dr. Roux, Paris, 75015, France
[4] AgroParisTech, F-75005 Paris, France

* co-senior authors

Abstract

Type II CRISPR-Cas systems introduce double strand breaks into DNA of invading genetic material and use DNA fragments to acquire novel spacers during adaptation. Double strand breaks are the substrate of several bacterial DNA repair pathways, paving the way for interactions between them and CRISPR-Cas systems. Here, we hypothesized that non-homologous end joining (NHEJ) interferes with type II CRISPR-Cas systems. We tested this idea by studying the patterns of co-occurrence of the two systems in bacterial genomes. We found that NHEJ and type II-A CRISPR-Cas systems only co-occur once among 5563 fully sequenced prokaryotic genomes. We investigated experimentally the possible molecular interactions causing this negative association using the NHEJ pathway from *Bacillus subtilis* and the type II-A CRISPR-Cas systems from *Streptococcus thermophilus* and *Streptococcus pyogenes*. Our results suggest that the NHEJ system has no effect on type II-A CRISPR-Cas interference and adaptation. On the other hand, we provide evidence for the inhibition of NHEJ repair by the Csn2 protein from type II-A CRISPR-Cas system. Our findings give insights on the complex interactions between CRISPR-Cas systems and repair mechanisms in bacteria and contribute to explain the scattered distribution of CRISPR-Cas systems in bacterial genomes.

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) arrays and their associated (Cas) proteins confer Bacteria and Archaea adaptive immunity against phages and other exogenous mobile genetic elements [1,2]. Yet, although most bacteria are infected by phages and other mobile genetic elements, CRISPR-Cas systems are absent from the majority of bacterial genomes[3,4]. The selective pressures and mechanisms that lead to the success of CRISPR-Cas systems in some clades and not others remain poorly understood.

CRISPR-Cas systems are classified into six types and twenty-seven subtypes, according to the Cas proteins they carry [3,5]. The recent development of CRISPR-Cas9-based genetic engineering technologies has made type II CRISPR-Cas systems the focus of many investigations. Type II systems include the CRISPR repeat-spacer array, three core genes (*cas1*, *cas2* and *cas9*), and a small trans-activating CRISPR RNA (tracrRNA) complementary to the CRISPR repeat sequence [6,7]. A fourth gene is involved in spacer acquisition, *csn2* in the type II-A [2,8–10], and *cas4* in type II-B systems [6]. A third subtype, type II-C, only requires *cas1*, *cas2* and *cas9* [3,6]. All the Cas proteins of type II systems are necessary for spacer acquisition [11,12], but only Cas9 is necessary for interference [13,14]. The Cas9 protein is guided by small CRISPR RNA (crRNA) to introduce double strand breaks (DSB) into target DNA[13,15]. A short conserved sequence (2-5bp) adjacent to the protospacer known as the PAM (protospacer Adjacent Motif) is essential to distinguish foreign from self DNA and can be different for CRISPR-Cas systems of the same type [16,17].

In bacteria, DSB can be repaired either by Homologous Recombination (HR) or by Non-Homologous End Joining (NHEJ). These mechanisms could thus affect the efficiency of CRISPR-Cas interference by repairing the breaks. Type II CRISPR-Cas systems introduce DSB at the same position in all copies of the target DNA molecule [18], and the concomitant lack of an intact DNA template should preclude the repair of these DSB by HR. However, NHEJ repairs DSB without requiring template DNA[19] and could mend DSB generated by Cas9. In Eukaryotic cells, breaks introduced by Cas9 can efficiently be repaired by NHEJ, a strategy now widely used to introduce indel mutations [20]. In bacteria, the NHEJ system requires two core proteins: Ku and a ligase [21]. Ligation is usually carried out by the LigD protein, but other ligases can be recruited by Ku when LigD is absent [19]. The system is complemented by additional proteins in certain cases [22]. Ku binds at the DSB and recruits the ligase to seal the break [23,24]. NHEJ offers a mean to repair DSB when only a single copy of the genome is available, such as after sporulation or during stationary phase [25,26]. NHEJ repair can be mutagenic [27], leading to up to 50% error rates in certain bacteria [24].

DNA repair pathways could also affect the acquisition of novel spacers by CRISPR-Cas systems because they modulate the availability of DSB and/or compete with the Cas machinery for the DNA substrate. Conversely, the action of Cas proteins at DSB could hinder DNA repair pathways. It was shown that novel spacers of type I CRISPR-Cas systems can be acquired after

DSB from RecBCD degradation products [28]. Importantly DNA repair pathways and CRISPR-Cas systems are composed of proteins with structural similarities and interacting with the same substrates [8]. For example, Cas4, a protein present in type I and type II-B systems shares structural and functional similarities with AddB[8,29] , a component of the AddAB repair pathway and a functional homolog of RecBCD[30]. In type II-A CRISPR-Cas systems, Csn2 binds and slides along free DNA ends in the same manner as the Ku protein of the NHEJ system [8]. Csn2 has been shown to be mandatory to acquire new spacers [2,11,12]. If Cas proteins and proteins involved in DNA repair mechanisms recognize the same substrate, a competition might arise leading to antagonistic interactions between the two processes.

The interaction between the NHEJ system and Cas9 is at the heart of the CRISPR-Cas based genetic engineering technologies, and we now search to understand it in bacteria. We hypothesized that the NHEJ system could interfere with the activities of type II CRISPR-Cas systems by repairing DSB generated by Cas9 during interference or by competing with Cas proteins for the same substrate during adaptation. Alternatively, type II CRISPR-Cas systems could interfere with NHEJ during repair. We tested these hypotheses by assessing the patterns of co-occurrence of the two systems in bacterial genomes. This revealed one single case of co-occurrence of both systems among 5563 bacterial genomes, suggesting strong negative interaction. We then studied experimentally the causes of this negative interaction, by introducing the NHEJ system from *B. subtilis* and/or the CRISPR-Cas system from *S. pyogenes* in *B. subtilis*, *S. thermophilus* and *S. aureus*.

Results

Negative association between NHEJ system and type II-A CRISPR-Cas systems

We detected CRISPR-Cas and NHEJ systems in 5563 fully sequenced bacterial genomes (Supplementary Table 1). The NHEJ pathway was present in 24.7% and the type II CRISPR-Cas system in 6.9% of the genomes, and these systems were very unevenly distributed among bacterial phyla (Supplementary Figure 1 and Supplementary Table 2). Firmicutes and Proteobacteria were the only phyla with genomes encoding enough type II CRISPR-Cas systems (resp. 209 and 101) and NHEJ (resp. 364 and 637), to perform robust statistical analyses (Supplementary Figure 1). A possible confounding factor when studying the distribution of bacterial defense and DNA repair pathways is that their abundance co-vary with genome size [31,32]. Accordingly, NHEJ systems were more frequent in larger genomes ($P<10^{-4}$, χ^2 test on a logistic fit). In contrast, type II CRISPR-Cas systems were only present in genomes smaller than 5Mb (Supplementary Figure 2). Hence, we focused our analysis on Firmicutes and Proteobacteria with genomes smaller than 5Mb. They represent 56.5% of the total number of genomes. In this sample, the size of the genomes with the NHEJ system was independent of the presence of a type II CRISPR-Cas system (P=0.99, Wilcoxon test).

We analyzed the patterns of co-occurrence of NHEJ and CRISPR-Cas systems to test if they were independently distributed. We observed that NHEJ and type II systems were negatively associated in Firmicutes ($P<10^{-4}$, Fisher Exact Test), but not in Proteobacteria (P=0.70, Fisher Exact Test) (Figure 1.b and Supplementary Figure 3). Note however that different subtypes of type II CRISPR-Cas systems are distributed differently in these two phyla. Proteobacteria encoded many type II-C and no type II-A systems, whereas Firmicutes encoded mostly type II-A systems (Figure 1.a). Type II-B systems were only detected in 9 genomes and will not be analyzed any further. To test if different subtypes could have different interactions with NHEJ systems, we looked at them separately. When studying co-occurrences of genes, it is important to consider that genomes are linked by a common evolutionary history, which decreases the degrees of the freedom of the statistical analyses. To check whether systems are negatively associated while taking phylogeny into account, we built a tree of Firmicutes and tested if the binary traits (presence of both systems) evolved independently using BayesTraits[33]. A strong negative association between NHEJ and type II-A CRISPR-Cas systems was observed (Bayes Factor BF=9.7, Figure 1.c), while no associations between NHEJ and type II-C CRISPR-Cas systems was detected. Only one genome among the 5563 encodes both NHEJ and type II-A: the actinobacteria *Eggerthella sp. YY7918*. In this genome, both NHEJ and type II-A systems seem intact, since the *cas* operon contains all four genes, lacking frameshifts or premature stop codons, and the adjacent CRISPR array encodes 44 spacers. We were also unable to detect anti-CRISPR proteins similar to the ones described in the literature[34–36].

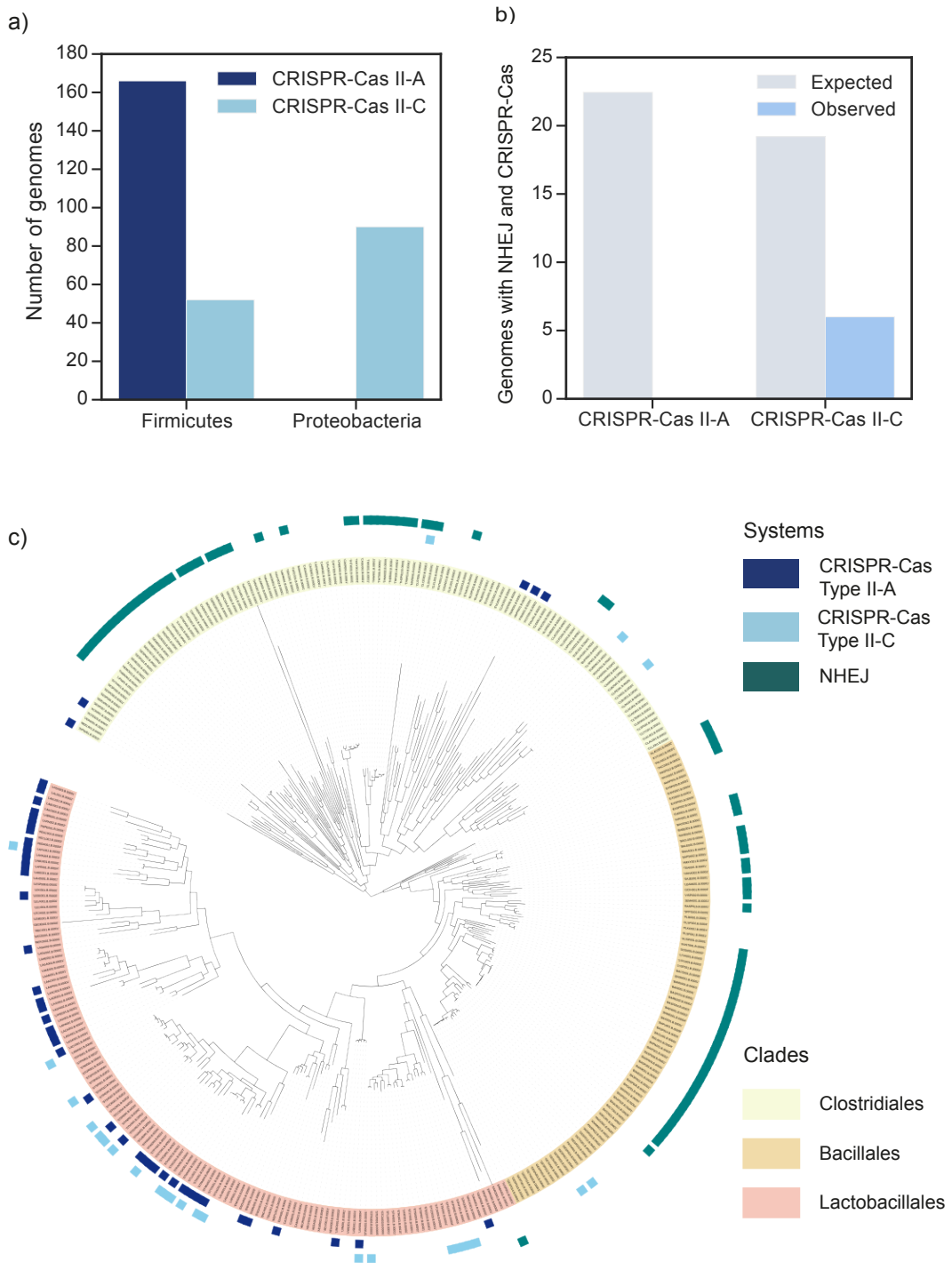


Figure 1: Negative association between NHEJ and type II-A CRISPR-Cas systems.
a, Distribution of the subtypes II-A and II-C in Proteobacteria and Firmicutes genomes. **b,** Associations between NHEJ and subtypes II-A and II-C CRISPR-Cas systems. Expected values correspond to the number of co-occurrences that would be obtained if the systems were randomly distributed. **c,** Presence of NHEJ and type II CRISPR-Cas systems in Firmicutes. A system is annotated as present in a given species when more than half of the genomes available for this species encode the system.

Taken together, these results show a strong negative association between NHEJ and type II-A CRISPR-Cas systems that is independent of the phylogenetic structure of the data. This negative association suggests the existence of a negative interaction between these systems in the bacterial cell. We devised three hypotheses to explain this negative association: 1) NHEJ impairs type II-A CRISPR-Cas interference, 2) NHEJ impairs type II-A CRISPR-Cas adaptation, *i.e.*, the ability of the system to acquire new spacers 3) type II-A CRISPR-Cas impairs NHEJ.

NHEJ system does not impact type II-A CRISPR-Cas interference

We first tested if the *B. subtilis* NHEJ system could affect type II-A CRISPR-Cas interference, using the previously described *S. aureus* model system[11]. The *ku* and *ligD* genes were cloned under the control of a Ptet promoter (plasmid pAB1) into *S. aureus* RN4220 cells. This system was able to circularize linearized plasmids after electroporation, showing it is functional (Supplementary Text 1, Supplementary Figure 4). The type II-A CRISPR-Cas system from *S. pyogenes* was introduced on plasmid pDB114 and programmed with a single spacer targeting phage phiNM4 (pMD021). *S. aureus* cells carrying both systems were then challenged in phage infection assays. A NHEJ system might facilitate phage escape from CRISPR-Cas by promoting the introduction of mutations at the target site through unfaithful repair, or by efficiently and faithful repairing DSB generated by Cas9, making CRISPR immunity inefficient.

First, the unfaithful repair of Cas9 breaks could lead to the formation of indels that would block further cleavages. The generation of such mutant phages should lead to a higher efficiency of plaquing (E.O.P) of phiNM4 when the NHEJ system is expressed. The CRISPR-Cas system provided a five order of magnitude reduction in the E.O.P. of phage phiNM4 when compared with a spacer-less control, and no significant increase in the number of plaques was observed upon NHEJ induction (Figure 2.a). To confirm that the small number of plaques obtained could not result from the unfaithful repair of Cas9 breaks through NHEJ, we sequenced the target position of 8 mutant phages. All mutants had a point mutation in the PAM and none presented an indel.

Second, the faithful repair of Cas9 breaks could lead to a cycle of repair and cleavage that would allow the production of functional phage particles. In this case it might not be possible to observe plaque formation as the competition between NHEJ and CRISPR interference might lower burst sizes. To test this hypothesis, we measured the efficiency of center of infection (E.C.O.I), *i.e.*, the number of cells that produce at least one functional phage particle after infection compared to the control strain (sensitive to the phage). One would expect higher E.C.O.I of phiNM4 when cells express the NHEJ system. The observed E.C.O.I was ~10^{-2} regardless whether the NHEJ system was induced or not (Figure 2.b).

We further tested whether NHEJ could reduce CRISPR-Cas9 immunity against plasmids. To this end, we cloned the PhiNM4 target sequence used above on plasmid pAB2 and transformed this plasmid in strains carrying the NHEJ system or not. While a control target-less plasmid could be efficiently introduced in the cells, no clones were recovered after transformation of pAB2 regardless of the presence of the NHEJ machinery. This shows that the CRISPR-Cas system efficiently blocks plasmid transformation and that the NHEJ system did not measurably reduce the efficiency of CRISPR immunity, nor introduced mutations in the target plasmid at a detectable rate (Figure 2.c).

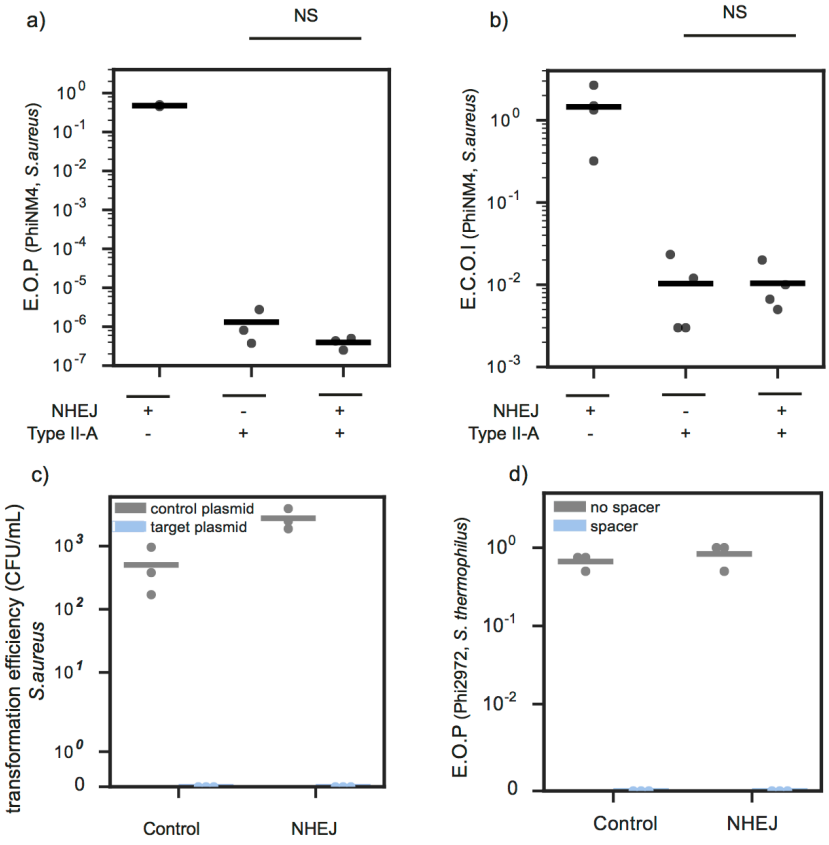


Figure 2: NHEJ system has no effect on type II-A CRISPR-Cas interference

a, Resistance to phage phiNM4 provided by the *S. pyogenes* CRISPR-Cas9 system in *S. aureus* in the presence (pAB1) or absence (pE194) of the NHEJ system from *B. subtilis* (n=3, mean, NS double sided t-test P=0.9999). **b,** Efficiency of Center Of Infection (E.C.O.I), *i.e.*, the proportion of cells that produce at least one functional phage particle, in the presence (pAB1) of absence (pE194) of the NHEJ system (n=4, mean, NS double sided t-test P=0.9998). **c,** Transformation efficiency of plasmid pT181 either empty or carrying a target sequence (pAB2) in *S. aureus* RN4220 cells expressing the CRISPR-Cas system from plasmid pMD021 in the presence (pAB1) of absence (pE194) of the NHEJ system from *B. subtilis* (n=3, mean). **d,** E.O.P. of phage Phi2972 on a bacteriophage insensitive mutant of *S. thermophilus* DGCC7710 carrying a spacer against Phi2972. Cells express either the *B. subtilis* NHEJ system from plasmid pAB66 or a control GFP from plasmid pAB69. (n=3, mean).

To confirm these results in a bacterium that naturally carries a type II-A CRISPR-Cas system, we measured interference against phage Phi2972 in *S. thermophilus*, in the presence or absence of the NHEJ system from *B. subtilis*. Genes *ku* and *ligD* were cloned under the control of a constitutive promoter on plasmid pNZ123 and introduced in a derivative of strain DGCC7710 whose CRISPR1 locus carries a spacer targeting phage Phi2972. The resistance provided by the CRISPR-Cas system was as strong in the presence of the NHEJ system as in the presence of a control GFP carried by the same plasmid (Figure 2.d). All in all, our results do not support the hypothesis that NHEJ affects type II-A CRISPR-Cas interference.

B. subtilis* NHEJ machinery does not prevent spacer acquisition in *S. aureus* and in *S. thermophilus

Ku and Csn2 bind the same type of substrate - linear double stranded DNA [8,37]- and might thus interfere antagonistically. To test if the NHEJ system affects spacer acquisition, we measured the cells' ability to acquire new spacers in presence of the NHEJ machinery. *S. aureus* cells carrying the NHEJ system (pAB1) and the type II-A CRISPR-Cas system (pRH87) were infected by phage PhiNM4 either with or without induction of the NHEJ system[11]. In this experiment, cells can escape phage infection either by capturing a novel spacer or by using other mechanisms of defense. Survivors were screened by PCR to check for acquisition of novel spacers and measure adaptation rate (Figure 3.a). No effect of the NHEJ system on the adaptation rate was observed. As a control the expression of Ku alone, ligD alone or GFP were also observed to have no effect (ANOVA, P=0,16) (Figure 3.b).

To corroborate these results, a similar experiment was performed in *S. thermophilus*. Cells carrying the *B. subtilis* NHEJ system or a control GFP on a plasmid were infected with phage Phi2972. We observed no difference the rate of novel spacer acquisition between cells expressing the NHEJ machinery or the GFP (Wilcoxon test, P=0.26) (Figure 3.c). Altogether these results indicate that NHEJ has no effect on the acquisition of novel spacers by a type II-A CRISPR-Cas system.

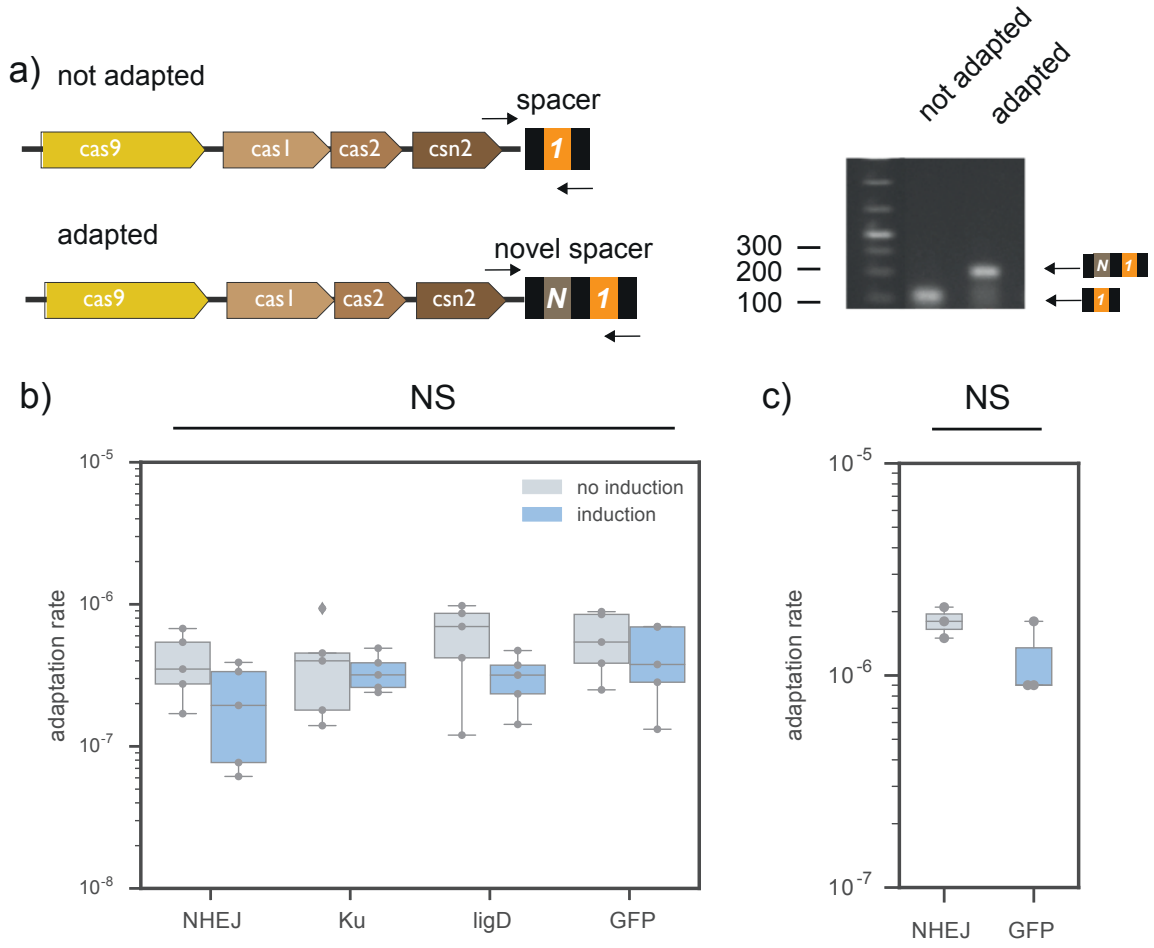


Figure 3 : NHEJ system does not impact adaptation of type II-A CRISPR-Cas system.
a, *S. aureus strain* RN4220 carrying the type II-A CRISPR-Cas system on plasmid pRH87 was challenged with phage phiNM4. Spacer acquisition was assessed by PCR on isolated colonies that survived the infection (oligonucleotides are depicted as black arrows). **b** Adaptation rate measured in the presence of NHEJ, ku, ligD or GFP carried by plasmids pAB23, pAB24, pAB25 and pAB62 respectively (n=5, ANOVA, NS P=0.5674). **c,** Adaptation rate of *S. thermophilus* DGCC7710 against phage Phi2972 when expressing the *B. subtilis* NHEJ system from plasmid pAB66 or a control GFP from plasmid pAB69 (n=3, two sided t test, NS P=0.91).

Csn2 inhibits NHEJ repair

As Csn2 binds to the same substrate as Ku, it could interfere with NHEJ repair[8,9,38]. To test this hypothesis, we reproduced the experiment that led to the discovery of the NHEJ system in *B. subtilis* [39]. When *B. subtilis* cells in stationary phase are irradiated by ionizing radiations (IR), the DSB generated are repaired by the NHEJ system, as other repair systems cannot function in those specific conditions. *B. subtilis* deleted for NHEJ do not survive irradiation as well as the wild-type. If type IIA CRISPR-Cas systems limit NHEJ repair, cells bearing a type IIA CRISPR-Cas system are expected to show increased sensitivity to irradiation.

B. subtilis cells expressing the type IIA CRISPR-Cas system from plasmid pRH087 were more sensitive to irradiation than cells carrying a control empty vector and showed the same level of sensitivity as the Δku-ligD mutant ($P<10^{-4}$, Wilcoxon, Figure 4.a). If the increased sensitivity provided by the CRISPR-Cas system is due to an impairment of NHEJ repair, then we expect to observe no cumulative effects when the NHEJ system is deleted and the CRISPR-Cas system added. Indeed, cells deleted for the NHEJ system and carrying the type II-A CRISPR-Cas system have the same survival as the ones deleted for the NHEJ system, pointing towards a interaction between the two systems. Another prediction that results from this hypothesis is that the CRISPR-Cas system should have no effect on the sensitivity to irradiation in species that lack a NHEJ system. To test this, we performed irradiation experiments on *S. aureus* cells carrying plasmid pRH87 or the control pC194. The presence or absence of the CRISPR-Cas system did not have an effect on survival in *S. aureus* (P=0.5, Wilcoxon, Supplementary Figure 5). Taken together, these results support the hypothesis that the type II-A CRISPR-Cas system impairs the NHEJ system.

To understand if a specific protein was responsible for this phenotype, we deleted or mutated individual *cas* genes from plasmid pRH87 and performed the same assay. While the effect size is small, the only mutant that significantly rescued *B. subtilis* cells upon irradiation was the delta *csn2* mutant (P=0.02, Student two sided t-test after validation of normality and homoscedasticity, Figure 4.b). When expressed alone, Csn2 was able to decrease survival of irradiated cells to the same level as the whole CRISPR-Cas system, while no effect could be observed with an empty vector or Cas9 alone ($P<10^{-4}$, Wilcoxon, Figure 4.c). In this set of experiments a possible concern is that Csn2 might be overexpressed which could lead to artifacts with no biological relevance. To prevent this issue, we expressed the whole *S. pyogenes* type II-A system or Csn2 alone from the natural promoter of the *cas* operon (plasmid pRH87 and pAB56 respectively). The expression of Csn2 in *B. subtilis* as measured by qPCR was 3.6-fold lower than the basal expression level of Csn2 in *S. pyogenes SF370* (Supplementary Text 2 and Supplementary Figure 6). This low level of expression might reflect what would happen after a natural horizontal gene transfer event.

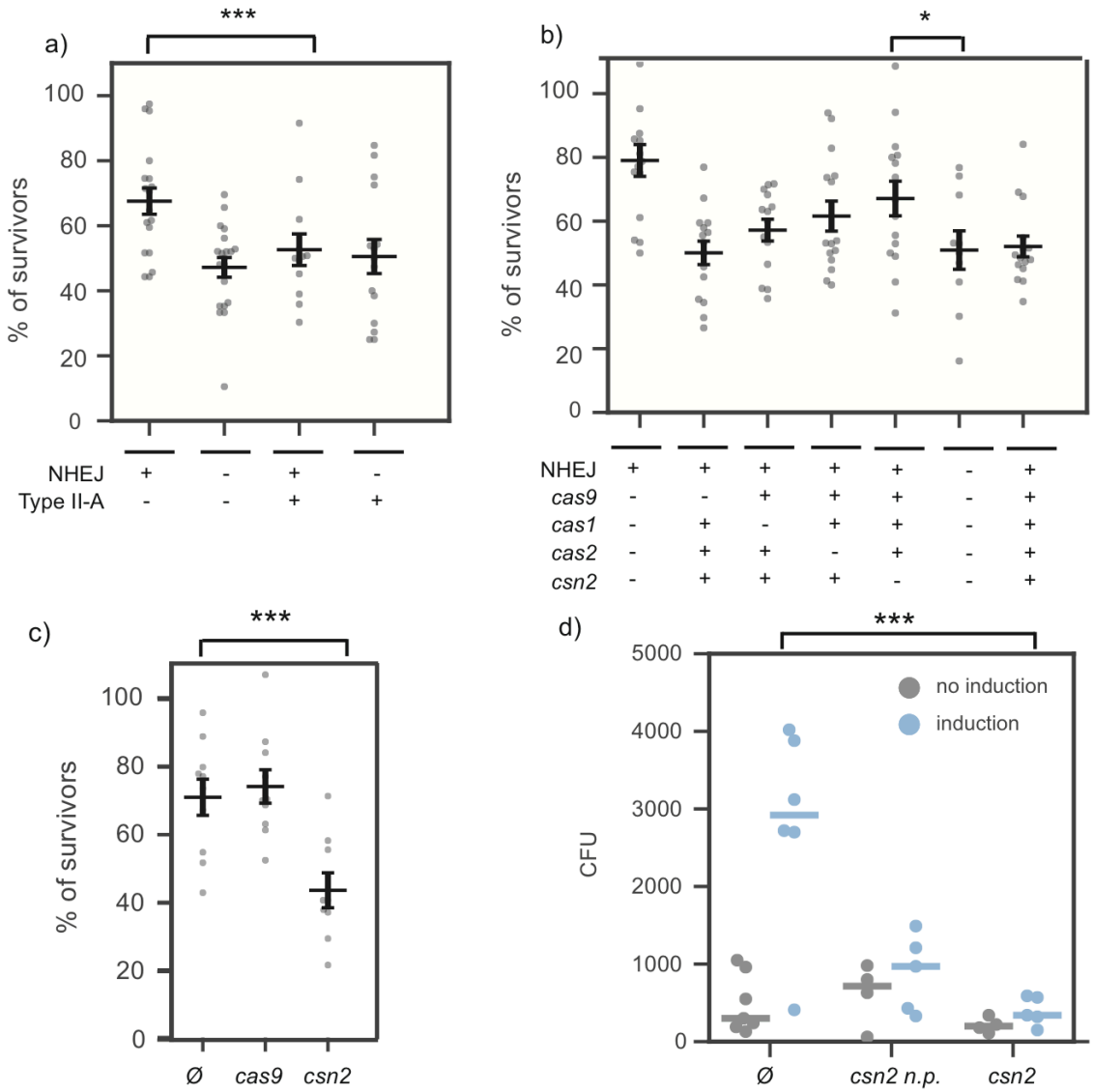


Figure 4: Type II-A CRISPR-Cas system impact NHEJ repair in *B. subtilis*

Survival rates of irradiated *B. subtilis* cells (a,b,c). Individual replicates (points) and average (horizontal bars) are shown. Error bars correspond to the standard error of the mean (s.e.m.). **a,** cells carrying the type IIA CRISPR-Cas system (pRH87) or the control empty vector (pC194), and deleted for *ku* and *ligD* or not (P=0.0009, Wilcoxon). **b,** *B. subtilis* carrying the CRISPR-Cas system with the dCas9 mutations (pRH121) or deleted for *csn2* (pRH63), *cas1* (pRH61), or *cas2* (pRH62)(P=0.02, Student two sided t-test). **c,** *B. subtilis* carrying the empty pC194 plasmid (Ø), expressing *csn2* from plasmid pAB56 or *cas9* from plasmid pDB114 (P=0.0048, Wilcoxon). **d,** A linearized plasmid providing resistance to chloramphenicol (pC194) was electroporated into *S. aureus* RN4220 cells carrying the NHEJ system either alone (plasmid pAB1, Ø) or with *csn2* cloned downstream of *ligD* (plasmid pAB81, csn2) or under the control of its natural promoter (plasmid pAB82, csn2 n.p.). The number of CFUs obtained with or without induction of the NHEJ system using aTc are reported. The number of CFU obtained without induction (grey bars) indicate the background of already circular DNA present in the sample before electroporation (P=0.0060, two sided t-test).

To obtain more direct evidence that Csn2 blocks NHEJ repair, we investigated its ability to inhibit the recircularization of linear plasmid DNA upon electroporation into *S. aureus*. The *csn2* gene was added to plasmid pAB1 which encodes Ku and LigD, either under the control of a Ptet promoter (pAB82), or under the control of the *cas* operon promoter (pAB81). We then electroporated a linearized plasmid providing resistance to chloramphenicol (pC194) into cells expressing the NHEJ system or both NHEJ and Csn2 (protocol presented in Supplementary Figure 4.a). The *B. subtilis* Ku and LigD were able to circularize the plasmid DNA in *S. aureus*, but we obtained on average 5-fold fewer colonies when Csn2 was co-expressed with Ku and LigD compared to the NHEJ system alone (Figure 4.d). In this assay the NHEJ system is strongly overexpressed compared to the natural expression of Ku and LigD in *B. subtilis* during stationary phase. Note that such overexpression was necessary to observe plasmid recircularization events in *S. aureus*. On the other hand, Csn2 was only slightly overexpressed compared to its expression level in *S. pyogenes SF370*. Altogether, these results show that Csn2 hinders NHEJ repair.

Discussion

We found that with the exception of a single case, NHEJ and type II-A CRISPR-Cas systems do not co-occur in fully sequenced bacterial genomes available to date. A possible incompatibility between NHEJ and type II-A CRISPR-Cas systems was investigated in a variety of experimental systems encompassing *S. aureus*, *B. subtillis* and *S. thermophilus*. Our results indicate that NHEJ does not affect CRISPR immunity against phages and plasmids, nor the capture of novel spacers. Previous studies showed that NHEJ repair pathways are able to repair Cas9-mediated DNA breaks in various bacterial species[18,40]. The efficiency of repair in these experimental setups was very low. Consistently, our results show that NHEJ repair cannot lead to a meaningful reduction in phage infectivity or plasmid transfer. Our results rather show that the Csn2 protein from type II-A CRISPR-Cas systems is able to inhibit NHEJ repair (Figure 5).

The strong avoidance of co-occurrences between NHEJ and type II-A systems was not observed with type II-C systems. This is consistent with the fact that type II-C systems lack Csn2. Csn2 is a multimeric toroidal protein that can bind double stranded DNA ends and slide inward through rotation-coupled translocation [8]. These DNA binding properties were noted in previous reports to be very similar to that of the Ku protein [8]. When present in the same cell, these two proteins will likely compete for the same substrate. We suggest that the binding of Csn2 at DNA-ends could block access to Ku or inhibit its function preventing efficient repair by the NHEJ machinery.

CRISPR-Cas systems are present in 47% of fully sequenced bacterial genomes[3] and this frequency might be much smaller in uncultivated bacteria [41]. This is in striking contrast with other defense systems, such as R-M systems, present on average at two copies per genome[31]. CRISPR-Cas systems are known to be transferred horizontally at a high rate[42], suggesting that they should spread in the bacterial world very rapidly if they were always advantageous. This brings to the fore the intriguing question of what is preventing further CRISPR rise in bacteria. Hypotheses that have been put forward include the cost of autoimmunity, the cost of limiting horizontal gene transfer, and the cost of inducible defenses [43–46]. Our results suggest another (non-mutually exclusive) reason: negative epistasis between the genetic background of a bacterium and a CRISPR-Cas system acquired by horizontal transfer can lead to a decreased fitness. In the present case, the type II-A CRISPR-Cas system affects the efficiency of NHEJ repair, thereby decreasing the fitness gain associated with the acquisition of the system. Note that type II-A systems are constitutively expressed in the bacteria where they have been studied (*S. pyogenes*[7], *S. thermophilus*[47]), and would thus likely also be expressed in the recipient upon horizontal gene transfer. We therefore propose that NHEJ is a barrier to the establishment of this type of CRISPR-Cas systems in bacteria.

We have observed an intriguing tendency of type II CRISPR-Cas systems to be absent from the largest genomes. DNA repair mechanisms are more frequent in larger genomes, presumably as a result of the presence of more abundant accessory functions[48], and to maintain constant genomic mutation rates[49]. If these larger genomes endure stronger selection for the presence of NHEJ, then incoming type II-A CRISPR-Cas systems will not be maintained in the genome. In agreement with the hypothesis of a trade-off between the two functions, nearly all of the largest genomes of Firmicutes encode NHEJ systems.

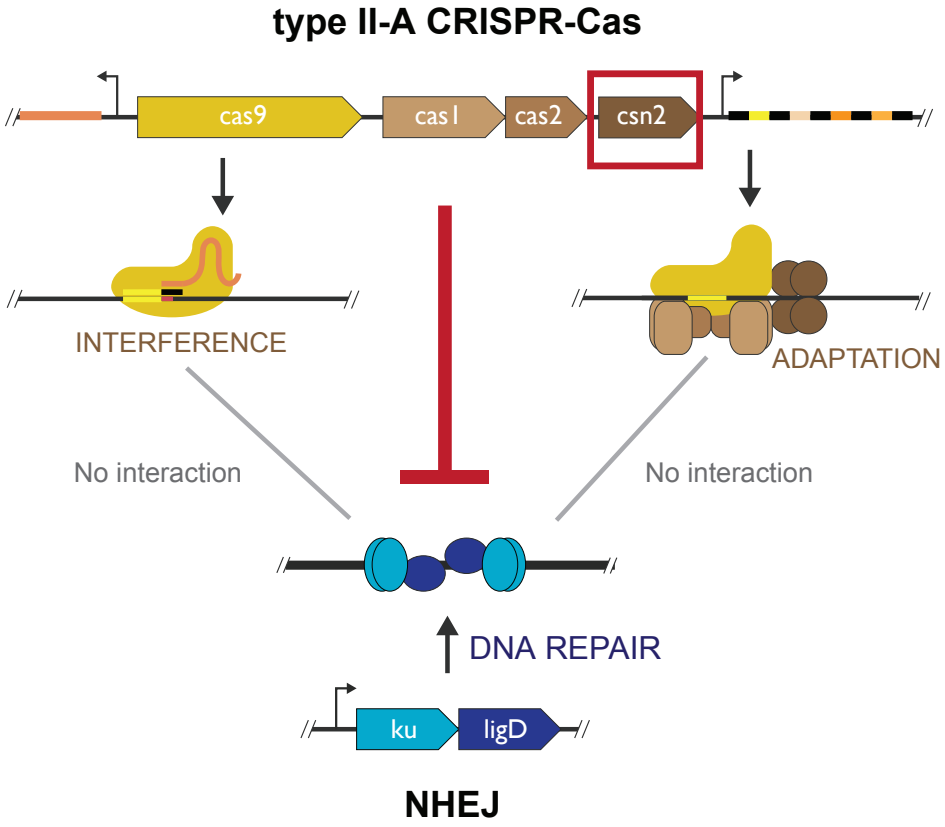


Figure 5: Graphical summary of the results. Three possible modes of negative interactions between type II-A CRISPR-Cas systems and NHEJ systems were tested: NHEJ could block CRISPR interference, NHEJ could block CRIPSR adaptation or CRISPR could block NHEJ repair. The last hypothesis was shown to be correct and Csn2 to be responsible for the inhibition of NHEJ repair.

Sorek and colleagues previously reported a positive effect of recBCD function on type I-E CRISPR spacer acquisition[28]. Since CRISPR-Cas systems acts by cutting DNA, interactions between these systems and DNA repair pathways might be numerous. These interactions are not only relevant to the evolution of bacterial genomes, but are also at the core of CRISPR genome editing technologies which rely on the repair of DNA breaks generated by Cas nucleases. In particular, the ability of Csn2 to block NHEJ repair could prove especially useful in genome editing experiments performed in Eukaryotes where NHEJ repair of Cas9-mediated breaks can compete with homology-directed repair and limit the efficiency with which precise modification are introduced.

Materials and Methods

Detection of repair systems and CRISPR-Cas systems

NHEJ and type II CRISPR-Cas systems were detected using MacSyFinder (default parameters)[4]. The published models were used for the detection of type II CRISPR-Cas systems[4]. To detect NHEJ, we retrieved protein profiles from TIGRFAM: Ku (PF02735), ligD (TIGR02777, TIGR02778, TIGR02779). We built a MacSyFinder model where the presence of Ku was defined as mandatory and that of LigD as accessory (Supplementary Text 2). Other ligases can indeed be recruited by Ku[19]. With this method, 74% of the systems detected encoded both Ku and LigD; 26% encoded only Ku. We compared these results to a previous analysis using other methods[30]. Only one out of 113 genomes was discordant (we identified a NHEJ system in *Sinorhizobium meliloti* were none had been found previously)[30].

Genome dataset

We analyzed 5563 complete genomes retrieved from NCBI RefSeq *(ftp://ftp.ncbi.nih.gov/genomes/,* last accessed in November 2016) representing 2437 species of Bacteria.

Phylogenetic analyses

We built persistent genomes for 245 Firmicutes genomes smaller than 5 Mb available in GenBank RefSeq (Dataset). The persistent genome of each clade was defined as the intersection of pairwise lists of orthologs that were present in at least 90% of the genomes. A list of orthologs was identified as reciprocal best hits using end-gap free global alignment, between the proteome of a pivot and each of the other strain's proteomes. *Bacillus subtilis* str.168 was used as pivot for each clade. Hits with less than 37% similarity in amino acid sequence and more than 20% difference in protein length were discarded. We made a persistent genome tree by concatenation of the multiple alignments of the persistent genes obtained with MAFFT v.7.205 (with default options, PMID: 23329690) and BMGE (with default options, PMID: 20626897). Missing genes were replaced by stretches of "-" in each multiple alignment. The tree was computed with IQ-TREE multicore v.1.5.4 under the LG+R10 model[50]. This model gave the lowest Bayesian Information Criterion (BIC) among all models available (option -m TEST in IQ-TREE). We made 1000 ultra-fast bootstraps to evaluate node support (options –bb 1000 –wbtl in IQ-TREE).
We applied BayesTraits v.2.0[33] to test the correlations among pairs of traits that adopt a finite number of discrete states. We ran two models (Independent and Dependent) in MCMC mode (priorAll exp 10) and computed the Bayes Factor BF which can be interpreted as follow : <2 weak evidence, >2 positive evidence, 5-10 strong evidence, >10 very strong evidence[51] .

Bacterial strains and growth conditions.

S. aureus strain RN4220 was grown in TSB (Tryptic Soy Broth) or TSA (Tryptic Soy Agar) at 37 °C. Whenever applicable, media were supplemented with chloramphenicol (10ug/ml), erythromycin (10ug/ml), tetracycline (100ng/mL), or spectinomycine (120ug/ml) to ensure the maintenance of pC194-derived, pE194-derived, pT181 and pLZ-derived plasmid respectively. Expression from ptet promoters was induced by addition of anhydrotetracycline (aTc) at 0.5ug/mL.
S. thermophilus strain DGCC7710 was grown in LM17 at 37 °C. Whenever applicable, media were supplemented with chloramphenicol (5ug/ml) to ensure the maintenance of pNZ123-derived plasmids.
B. subtilis strain 168 was grown in LB or LB agar at 37 °C. Whenever applicable, media were supplemented with chloramphenicol (5ug/ml) or erythromycin (1ug/ml) to ensure the maintenance of pC194-derived plasmids and the integration of pMUTIN4-derived plasmids.

Plasmids and strains construction
The cloning strategies employed for each plasmid are summarized in Supplementary Table 3 and the primers used are listed in Supplementary Table 4. PCR fragments were assembled using Gibson assembly[52] unless mentioned otherwise. Plasmids pAB2, pAB17, pAB18, pAB56 were obtained by PCR followed by blunt end ligation. Plasmid pMD021 was assembled by Golden Gate[53].

CRISPR-Cas interference efficiency assay using phages
We used two types of assays to assess the impact of Ku and LigD on CRISPR-Cas immunity. **Phage titre assay**. Top agar lawns supplemented with 5mM CaCl2 and inoculated with strains bearing the NHEJ system or not were poured on selective plates (with aTc for induction in *S. aureus*). We spotted serial dilutions of PhiNM4 or Phi2972 on the lawns of *S. aureus* and *S. thermophilus* respectively. *S. aureus* strain RN4220 carried the *S. pyogenes* CRISPR-Cas system on plasmid pDB114 or a derivative with spacer `5′-AAAATGTTTTAACACCTATTAACGTAGTAT-3′` (pMD021). *S. thermophilus* strain DGCC7710 and a bacteriophage insensitive mutant of strain DGCC7710 carrying spacer `5′-TGTTAAAAGAAGCACTAGAGGTGATTTACG-3′` in the first position of the CRISPR-1 locus were used. E.O.P was determined after overnight incubation at 37°C. **Productive infection assays**. Cells were diluted 1:100 from overnight cultures in TSB supplemented with 5mM CaCl2 and the appropriate antibiotics., and incubated at 37°C. The NHEJ system was induced using aTc at OD600 0.2. After 30 minutes of incubation allowing the expression of the NHEJ system, we added phage PhiNM4 at a M.O.I (Multiplicity of Infection) of 1. Adsorption was allowed for 5 minutes at 37°C with shaking. Cells were then put on ice and washed twice with ice cold TSB. We then diluted and spotted them on top agar lawns of RN4220 supplemented with CaCl2. E.C.O.I was determined after overnight incubation at 37°C.

CRISPR-Cas interference efficiency assay using plasmids

Cells carrying a type II-A CRISPR-Cas systems (pRH87) and the NHEJ system (pAB1) or the empty vector as a control (pE194) were made electro-competent as follow: cells were grown until OD 0.4, induced by adding aTc and further grown to OD 0.8. Cells were then washed twice with ice-cold water, once with 10% glycerol and resuspended in 1/100 of their volume in 10% glycerol. 100 ng of plasmid pT181 or pAB2 were electroporated in 50ul of competent cells (2500V, 25µF, 100Ω and 2mm cuvettes). Cells were then incubated in 1ml TSB for one hour at 37°C and plated on tetracycline only. Transformation efficiency was assessed after overnight incubation at 37°C.

Adaptation assays

The spacer acquisition assay was described elsewhere[11]. We mixed cells from overnight cultures (induced or non-induced) with phage (M.O.I value of 1) in top agar supplemented with 5mM CaCl2 and poured them on plates containing appropriate antibiotics and supplemented with aTc when necessary, followed by overnight incubation at 37°C. For *S. aureus,* single colonies were resuspended in lysis buffer (250mM KCl, 5mM $MgCl_2$, 50mM Tris-HCl at pH 9.0, 0.5% Triton X-100) supplemented with 20ng/mL lysostaphin and incubated at 37°C for 10 min, then 98°C for 10 min. Following centrifugation (11 000g), 1ul of the supernatant was used as template for DreamTaq PCR amplification with primers AB23 and AB24. We provide a list of 15 acquired spacers in supplementary table 5. For *S. thermophilus*, single colonies were resuspended in 10ul of water, 1ul of which was used as template for DreamTaqPCR amplification with primers AB103 and AB104. The PCR reactions were analyzed on 2% agarose gels. Adaptation rates were computed as the estimated number of clones that acquired a spacer divided by the estimated number of cells in the initial population.

Irradiation assay

The NHEJ repair assay was described elsewhere[39]. 100ul of overnight cultures of *B. subtilis* were irradiated at 100 Gy (RS Xstrahl, 42 minutes, 250kV, 12mA, 30cm from focal point). We plated 1:10 000 dilutions on appropriate antibiotics. CFUs were determined after overnight incubation at 37°C. Survival rates were determined as the ratios of CFUs obtained for irradiated cells over CFUs obtained for non-irradiated cells.

References

1. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).
2. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science.* **315,** 1709–12 (2007).
3. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR Cas systems. *Nat. Rev. Microbiol.* **13,** 722–736 (2015).
4. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One* **9,** e110726 (2014).
5. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR Cas systems. *Nat. Rev. Microbiol.* (2017). doi:10.1038/nrmicro.2016.184
6. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* **42,** 6091–6105 (2014).
7. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
8. Arslan, Z. *et al.* Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.* **41,** 6347–6359 (2013).
9. Ellinger, P. *et al.* The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. *J. Struct. Biol.* **178,** 350–362 (2012).
10. Lee, K. H. *et al.* Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins Struct. Funct. Bioinforma.* **80,** 2573–2582 (2012).
11. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR Cas adaptation. *Nature* **519,** 199–202 (2015).
12. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR – Cas adaptation. *Genes Dev.* **29,** 356–361 (2015).
13. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science.* **337,** 816–21 (2012).
14. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR / Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39,** 9275–9282 (2011).
15. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2579-86 (2012).
16. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1390–1400 (2008).
17. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155,** 733–740 (2009).
18. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res.* **44,** 4243–4251 (2016).
19. Shuman, S. & Glickman, M. S. Bacterial DNA repair by non-homologous end joining.

Nat. Rev. Microbiol. **5,** 852–61 (2007).
20. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).
21. Della, M. *et al.* Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science.* **306,** 683–685 (2004).
22. Aniukwu, J., Glickman, M. S. & Shuman, S. The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. *Genes Dev.* **22,** 512–527 (2008).
23. Bowater, R. & Doherty, A. J. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genet.* **2,** e8 (2006).
24. Gong, C. *et al.* Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. *Nat. Struct. Mol. Biol.* **12,** 304–312 (2005).
25. Pitcher, R. S. *et al.* NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation. *DNA Repair (Amst).* **6,** 1271–1276 (2007).
26. Moeller, R. *et al.* Role of DNA repair by nonhomologous-end joining in Bacillus subtilis spore resistance to extreme dryness, mono- and polychromatic UV, and ionizing radiation. *J. Bacteriol.* **189,** 3306–3311 (2007).
27. Paris, Ü. *et al.* NHEJ enzymes LigD and Ku participate in stationary-phase mutagenesis in Pseudomonas putida. *DNA Repair (Amst).* **31,** 11–18 (2015).
28. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520,** 505–510 (2015).
29. Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR Associated Protein Cas4 Is a 5' to 3' DNA Exonuclease with an Iron-Sulfur Cluster. *PLoS One* **7,** e47232 (2012).
30. Rocha, E. P. C., Cornet, E. & Michel, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet.* **1,** e15 (2005).
31. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42,** 10618–10632 (2014).
32. Silva, F. J., Latorre, A. & Moya, A. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet.* **19,** 172–176 (2003).
33. Pagel, M. & Meade, A. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible Jump Markov Chain Monte Carlo. *Am. Nat.* **167,** 808–825 (2013).
34. Rauch, B. J. *et al.* Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell* **168,** 150–158 (2016).
35. Pawluk, A. *et al.* Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.* **1,** 16085 (2016).
36. Hynes, A. P. *et al.* An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. *Nat. Microbiol.* **2,** 1374–1380 (2017).
37. Nam, K. H., Kurinov, I. & Ke, A. Crystal structure of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. *J. Biol. Chem.* **286,** 30759–30768 (2011).
38. Koo, Y., Jung, D. K. & Bae, E. Crystal structure of streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS One* **7,** e33401 (2012).
39. Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in

bacteria. *Science.* **297,** 1686–1689 (2002).
40. Su, T. *et al.* A CRISPR-Cas9 Assisted Non-Homologous End-Joining Strategy for One-step Engineering of Bacterial Genome. *Sci. Rep.* **6,** 37895 (2016).
41. Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defense systems. *Nat. Commun.* **7,** 10613 (2016).
42. Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62,** 718–29 (2006).
43. Bondy-Denomy, J. & Davidson, A. R. To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends Microbiol.* **22,** 218–25 (2014).
44. Jiang, W. *et al.* Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet.* **9,** e1003844 (2013).
45. Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. a. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* **12,** 177–86 (2012).
46. Westra, E. R. *et al.* Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr. Biol.* **25,** 1043–1049 (2015).
47. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
48. Nimwegen, E. van. Scaling laws in the functional content of genomes. *Trends Genet.* **19,** 476–479 (2003).
49. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148,** 1667–1686 (1998).
50. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
51. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. Introducing Markov Chain Monte Carlo. *Markov Chain Monte Carlo in Practice* 512 (1996). doi:10.1007/978-1-4899-4485-6_1
52. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–5 (2009).
53. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3,** (2008).

Acknowledgements
We thank Philippe Horvath for providing *S. thermophillus* strain DGCC7710 and phage Phi2972 and Sylvain Moineau for providing plasmid pNZ123. We also want to thank them both for their very helpful advice. We thank Matt Deyell for plasmid pMD021. A.B. is a member of the "Ecole Doctorale Frontière du vivant (FdV)". Funding: This work was supported by the European Research Council (ERC) under the Europe Union's Horizon 2020 research and innovation programme (grant agreement No [677823]); the French Government's Investissement d'Avenir program; Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' [ANR-10-LABX-62-IBEID]; the Pasteur-Weizmann consortium.

Author contributions.
A.B., E.R., M.T. and D.B. designed the research. A.B., A.C.V., C.B, L.C.. performed the experiments. A.B., E.R., M.T. and D.B wrote the manuscript.

Competing financial interests.
The authors declare no competing financial interests.

Supplementary Materials

Supplementary Text 1: Functionality of NHEJ system from *B. subtilis in S. aureus*

NHEJ system from *B. subtilis* is functional in *S. aureus*

We first tested the functionality of the NHEJ system from *B. subtilis* in *S. aureus*. To do so, we cloned *ku* and *ligD* from *B. subtilis* into the staphylococcal vector pE194 under a tetracycline-inducible promoter Ptet (plasmid pAB1) and introduced it in RN4220 *S. aureus* cells. We then electroporated a linearized plasmid providing resistance to chloramphenicol (plasmid pC194) either in the presence or absence of anhydrotetracycline (aTc). Only cells that re-circularized the plasmid can form colonies on chloramphenicol plates (Supplementary Figure 4.a). We obtained 5 times more colonies when the NHEJ system was induced compared to the uninduced control (Supplementary Figure 4.b). Colonies were checked by sequencing the junction, which showed repair patterns typical of NHEJ (Supplementary Figure 4.c). We therefore concluded that the NHEJ system from *B. subtilis* was functional in *S. aureus.*

NHEJ Functionality Assay

The plasmid pC194 was linearized by PCR using primers B329 and B330 (Supplementary Table 4). Strains with the plasmids carrying the NHEJ system were grown to OD 0.3 and the NHEJ system was induced by adding aTc. Cells were grown to OD 0.8 and made electro-competent by washing them three times in ice-cold water, supplemented with 10% glycerol for the last wash, and concentrated 100 fold. We transformed 200ug of linearized pC194 in those electro-competent cells and added aTc to the recovery medium. Cells were plated on selective media and incubated overnight at 37°C. We resuspended single colonies in lysis buffer with 15ng/mL lysostaphin and incubated them at 37°C for 10 min, then 98°C for 10 min. Following centrifugation (11 000g), 1 ul of the supernatant was used as template for DreamTaqPCR amplification with primer A9, A10 (Supplementary Table 4). PCR products were then purified and sequenced.

Supplementary Text 2: Expression of NHEJ and Csn2 in strains used in the study

RNA extraction

RNA was extracted from strains *B. subtilis* 168, *B. subtilis* 168 + pRH87, *B. subtilis* 168 + pAB56, *S. pyogenes* SF370, *S. aureus* RN4220 + pAB82, *S. aureus* RN4220 + pAB1 + pRH87. Overnight cultures were diluted 1:100 in 2ml and incubated at 37°C for 3 hours. For strains with pAB1 or pAB82 plasmids, aTc (0.5 ug/ul) was added after 1 hour of incubation. 4 ml of RNAprotect bacteria reagent (Qiagen) were added to the cultures, which were then vortexed briefly and incubated at room temperature for 5 minutes. The tubes were spinned down at 4000 g for 5 minutes. Cell pellets of *B. subtilis* and *S. pyogenes* were resuspended in 200 ul of lysozyme buffer (lysozyme 20 mg/ml). *S. aureus* cell pellets were resuspended in 200ul of lysostaphin

solution (lysostaphin 5mg/mL). After 1 hour incubation at 37 °C, 1 ml of trizol was added, and regular trizol reagent procedures for purifying the total RNA were followed.

RT-qPCR

All the RNA samples were treated with DNase (Turbo DNase free kit, Ambion), then all the RNA samples (1 ug for each sample) were reverse transcribed into cDNA using the Transcriptor First strand cDNA synthesis Kit (Roche). The qPCR was performed using 1 ul of the reverse transcription reaction and the Faststart essential DNA green master mix (Roche) in a LightCycle 96 (Roche). Probes and PCR primers are listed below. Relative gene expression was computed using the ΔΔCq method ($2^{Cq_{TAR}-Cq_{REF}}$) where Cq_{REF} is the quantification cycle value for the 16s rRNA and Cq_{TAR} for the tested gene. Data is shown relative to expression in the wild-type strain (Ku in *B. subtilis* 168 or Csn2 in *S. pyogenes* SF370).

Targeted genes	Primer name	Sequences (5' to 3')
Ku	LC1340_Ku_For	GGATCGATCAGCTTCGGATTAG
Ku	LC1341_Ku_Rev	TGGTGCGTGATCCTCTTTATG
Csn2	LC1342_csn2_For	GCAAACTCCGATGAAAGACTTG
Csn2	LC1343_csn2_Rev	ACCGCCTCTTAATGGAATCG
16s_rRNA	LC1344_16s_For	AGGCAGCAGTAGGGAATCTT
16s_rRNA	LC1345_16s_Rev	GCTGCTGGCACGTAGTTAG

Supplementary Text 3 : Model for NHEJ system detection with MacsyFinder

```
<system inter_gene_max_space="5" min_mandatory_genes_required="1"
min_genes_required="1">
<gene name="ku" presence="mandatory" loner="1"/>
<gene name="ligD1" presence="accessory"/>
<gene name="ligD2" presence="accessory"/>
<gene name="ligD3" presence="accessory"/>
</system>
```

Supplementary Table 1 : Detection of Type II CRISPR-Cas systems and NHEJ in Bacteria and Archea

Csv File

Supplementary Table 2 : Detected systems and their frequencies by genomes

Systems	Bacterial Genomes (2482)	Frequency (%) (Bacteria)
Type I CRISPR-Cas	1584	28.5
Type II CRISPR-Cas	384	6.9
Type III CRISPR-Cas	306	5.5
NHEJ	1376	24.7

Supplementary Table 3: Plasmids used in this study

Plasmids from other studies		**Described in**
pC194/pE194	Replicative plasmid in *S. aureus*, respectively CmR, ErmR,	[1,2]
pT181	Replicative plasmid in *S. aureus,* tetR	[3]
pMUTIN4	Integrative plasmid in *B. subtilis*, ErmR	[4]
pNZ123	Replicative plasmid in *S. thermophilus*, CmR	[5]
pLZ12	Replicative plasmid in *S. aureus* and *E. coli*	[6]
pCN57	Plasmid with a working GFP in *S. aureus*	[7]
pRH87	pC194 with type II-A CRISPR-Cas system from *S. pyogenes* with one spacer in the CRISPR array	[8]
pRH61,pRH62,pRH63	same as pRH87 but without respectively *cas1,cas2*, *csn2*	[8]
pRH121	same as pRH87 but with *dcas9* instead of *cas9*	[8]
pDB114	pC194 with *cas9* and reprogrammable spacer	[9]
pWJ153	pE194 Ptet inducible target vector	[9]

Plasmids created for this study		Insert (Primer ; Template)		Vector (Primer ; Template)		Cloned in
pAB1	pE194 with NHEJ system from *B. subtilis* under tet promoter (Ptet)	AB+AB4	*B. subtilis* genomic DNA	AB1+AB2	pWJ153	*S. aureus*
pAB2	pT181 with a protospacer from PhiNM4 phage			AB9+AB10	pT181	*S. aureus*
pAB12	pMUTIN with ykoV homologies (to build ykoV mutant in *B. subtilis*)	AB69+AB70	*B. subtilis* genomic DNA	AB67+AB68	pMUTIN 4	*E. coli*
pAB13	pMUTIN with ykoU homologies (to build ykoU mutant in *B. subtilis*)	AB73+AB74	*B. subtilis* genomic DNA	AB71+AB72	pMUTIN 4	*E. coli*
pAB17	pE194 with ligD under tet promoter (Ptet)			AB85+AB86	pAB1	*S. aureus*
pAB18	pE194 with Ku under tet promoter (Ptet)			AB87+AB88	pAB1	*S. aureus*
pAB23	pLZ12 with NHEJ under tet promoter (Ptet)	AB95+AB96	pAB1	AB97+AB98	pLZ12	*E. coli*
pAB24	pLZ12 with Ku under tet promoter (Ptet)	AB95+AB96	pAB18	AB97+AB98	pLZ12	*E. coli*
pAB25	pLZ12 with ligD under tet promoter (Ptet)	AB95+AB96	pAB17	AB97+AB98	pLZ12	*E. coli*
pAB62	pLZ12 with GFP under tet promoter (Ptet)	ACV69+ACV70	pCN57	ACV64+ACV65 ACV67+ACV68	pAB23	*E. coli*
pAB66	pNZ123 with NHEJ under P8 constitutive *S. thermophilus* promoter	AB244+AB245 AB246+AB247	pAB1 ; *S. thermophilus* DNA	AB242+AB243	pNZ123	*E. coli*
pAB69	pNZ123 with GFP under P8 constitutive *S. thermophilus* promoter	AB50+AB251 AB252+AB253	pCN57 ; *S. thermophilus* DNA	AB248+AB249	pNZ123	*E. coli*
pAB56	pC194 with csn2 under type II-A natural promoter	Reciculrized		AB224+AB225	pRH87	*S. aureus*
pAB81	pAB1 with csn2 cloned directly after ligD	AB280+AB281	pAB56	AB278+AB279	pAB1	*S. aureus*
pAB82	pAB1 with csn2 cloned with the promoter of the cas operon of *S. pyogenes*	AB284+AB285	pAB56	AB282+AB283	pAB1	*S. aureus*
pMD021	pC194 with *cas9* and spacer targeting PhiNM4	D035+D036	Hybridized primers	Digested with Bsa1	pDB114	*S. aureus*

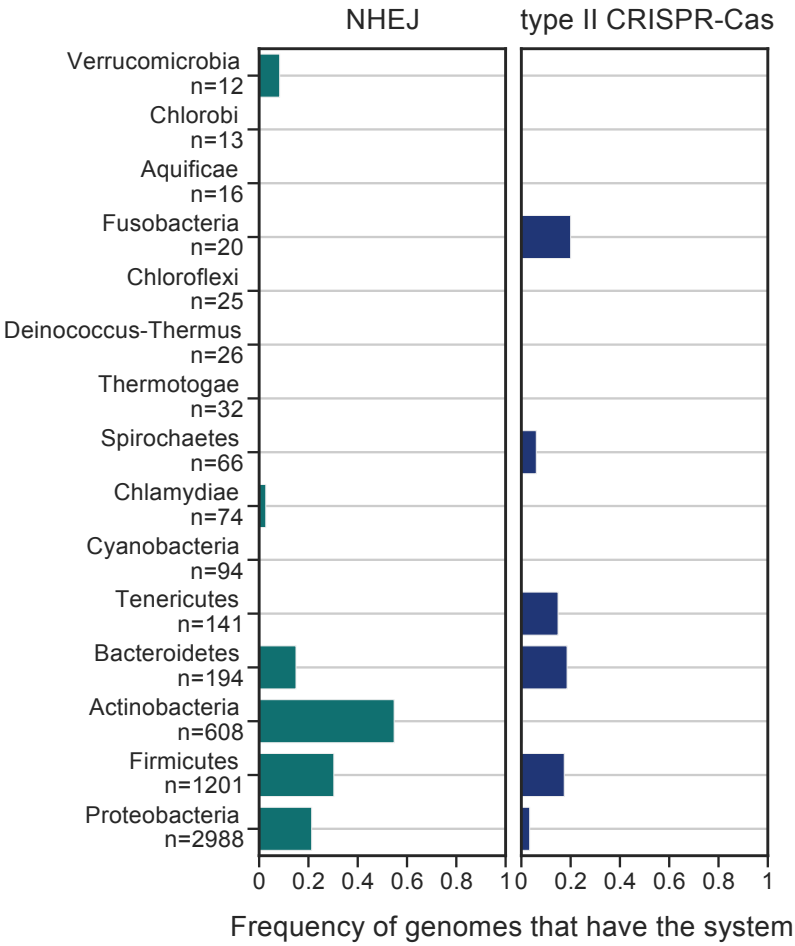
Supplementary Table 4: Primers used in the study

AB1	GCTCTTTCCGTGGTTCATATTTATCAGAGCTCGTGCTATAAT
AB2	GAGTACGATTCATTTGATATGCCTCCTCTAGGTCATTTGATATGCCTCCG
AB3	CAAATGACCTAGAGGAGGCATATCAAATGAATCGTACTCCTTCTCTTCAC
AB4	GATAAATATGAACCACGGAAAGAGCTGACTTCATTAG
AB9	AAAAATGTTTTAACACCTATTAACGTAGTATGGAGTGGCTAGCATTTTGC
AB10	CCATACTACGTTAATAGGTGTTAAAACATTTTTAAACTGCTTTTCAGAAC
AB67	AGTCCCAGGTGTACCTGATCAGAAAACCGCCTCGCG
AB68	ACAAAGTCAACAATCTCAACGATTCTCCGTGGGAACAAACGG
AB69	GTTTGTTCCCACGGAGAATCGTTGAGATTGTTGACTTTGTTCAGCTTCAG
AB70	TCACCGCGAGGCGGTTTTCTGATCAGGTACACCTGGGACTTGAG
AB71	ACTATGCGCCGGATTTTGTTAGAAAACCGCCTCGC
AB72	TGTTAAGTATTGAACAGCTGGATTCTCCGTGGGAACAAACG
AB73	GTTTGTTCCCACGGAGAATCCAGCTGTTCAATACTTAACAATTCTCCAAG
AB74	TCACCGCGAGGCGGTTTTCTAACAAAATCCGGCGCATAGTCC
AB85	CGCATGGCGTTTACCATGCA
AB86	GGTCATTTGATATGCCTCCGG
AB87	GAAGTCAGCTCTTTCCGTGGTTC
AB88	GCGCTATGATGTGCCGGAG
AB95	CTAATGAATTCATCTGCAGGAAAGAAATTAGATAAATCTCTCATATCTTTTATTCAATAATCGCAT
AB96	GGTCGTCAGACTGATGGGCCAATTATAGCACGAGCTCTGATAAATATGAACCAC
AB97	TCAGAGCTCGTGCTATAATTGGCCCATCAGTCTGACGAC
AB98	GAGATTTATCTAATTTCTTTCCTGCAGATGAATTCATTAGGATCCAGA
AB224	ATTTACATGGTGAAAGAAATAATTGTATTGCAAACTCC
AB225	TTGCCTCCTAAAATAAAAAGTTTAAATTAAATCC
AB242	ACTTCGAACTAGCAATACTGCTCTCTAGAGAATTCAGTACTGGATCT
AB243	TTCCGTGGTTCATATTTATCCTCAAGCTTCTCGAGTGCATATTTTCG
AB244	CTCGTGCGAGGTTTTTACATATGAATCGTACTCCTTCTCTTCACACTAAAG
AB245	ATGCACTCGAGAAGCTTGAGGATAAATATGAACCACGGAAAGAGCTGAC
AB246	GTACTGAATTCTCTAGAGAGCAGTATTGCTAGTTCGAAGTCATCCTTTTTTATAGG
AB247	AGAGAAGGAGTACGATTCATATGTAAAAACCTCGCACGAGTAGTTATTT
AB248	CGAACTAGCAATACTGTAAGAGCTCTCTAGAGAATTCAGTACTGGATCT
AB249	CAAATAAGGCGCGCCTATTCCAAGCTTCTCGAGTGCATATTTTCG
AB250	CTCGTGCGAGGTTTTTACATATGAGTAAAGGAGAAGAACTTTTCACTGGAG
AB251	ATATGCACTCGAGAAGCTTGGAATAGGCGCGCCTTATTTGTATAGTT
AB252	ACTGAATTCTCTAGAGAGCTCTTACAGTATTGCTAGTTCGAAGTCATCCT
AB253	AGTTCTTCTCCTTTACTCATATGTAAAAACCTCGCACGAGTAGTTATTT
AB278	GGTATAATACTCTTAATAAAAAGTCAGCTCTTTCCGTGGT
AB279	GTAAATTTGCCTCCTAAAATCATTAGTCAGCTCTTTTTCTTCAACTGATG

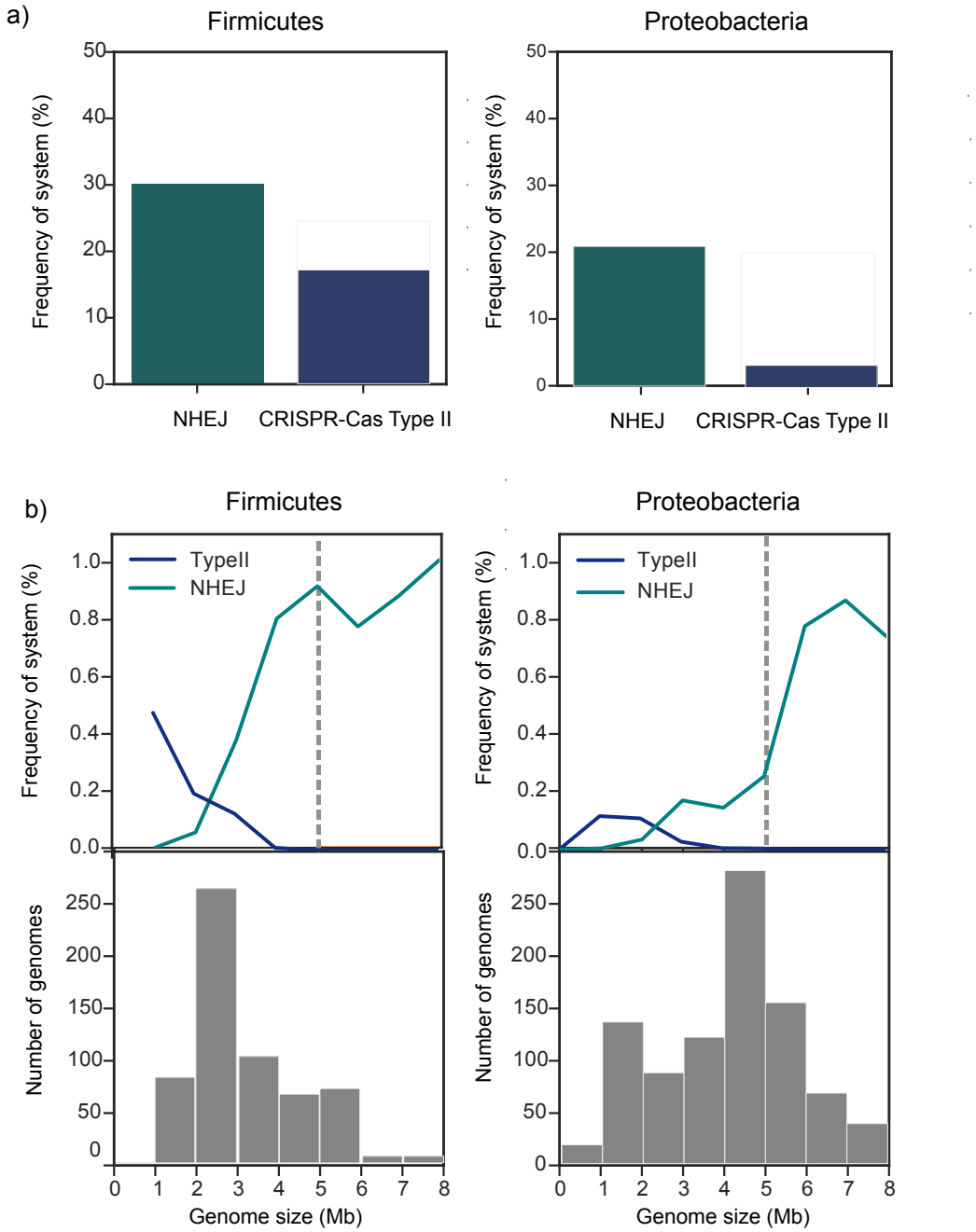
AB280	AGAAAAAGAGCTGACTAATGATTTTAGGAGGCAAATTTACATGGTGAAAGAA
AB281	ACCACGGAAAGAGCTGACTTTTTATTAAGAGTATTATACCATATTTTTAGTTATTAAG
AB282	AAATGCAGTAATACAGGGGCTCGTGCTATAATTATACTAATTTTATAAGGAGG
AB283	CAAAAAATATTACCCAATACGCTCTGATAAATATGAACCACGGAAAGAG
AB284	TGGTTCATATTTATCAGAGCGTATTGGGTAATATTTTTTGAAGAGATATTTTGAAAAAG
AB285	TTAGTATAATTATAGCACGAGCCCCTGTATTACTGCATTTATTAAGAG
ACV64	ACACATGGCATGGATGAACTATACAAATAATTTCTAAATAAGAATATTTGGAGAGCACCGTTC
ACV65	AGCATAACCTTTTTCCGTGATGGTA
ACV66	TTGATATGCCTCCTCTAGGTCATTTG
ACV67	TATAAATTTAACGATCACTCGTTACCATCACGG
ACV68	TCCAGTGAAAAGTTCTTCTCCTTTACTCATTTGATATGCCTCCTCTAGGTCATTTG
ACV69	ATGAGTAAAGGAGAAGAACTTTTCACTGGA
ACV70	TTATTTGTATAGTTCATCCATGCCATGTGTAAT
D035	AAAAATGTTTTAACACCTATTAACGTAGTATGTTTTAGAGCTATGCTGTTTTGA
D036	ATACTACGTTAATAGGTGTTAAAACATTTTTGTTTTGGGACCATTCAAAACAGC
B329	ACACTGAGACTTGTTGAGTTTGCCTAAAAACCTACAGAAG
B330	CTCCACAGGATGATTTCGTAAAACTATATGATTTACCCCTAAATCT
A9	TCAACGCACAATAAATTTTCTCGGC
A10	TACTTAAAAGAAATTGATCCAACCG
AB103	GCCCTCGAGTTGACAAGGACAGTTATTG
AB104	CAATTCGAATCTTGATTTGCTGTC

Supplementary Table 5 : Example of spacers acquired during adaptation experiments of *S. aureus* against the phage PhiNM4

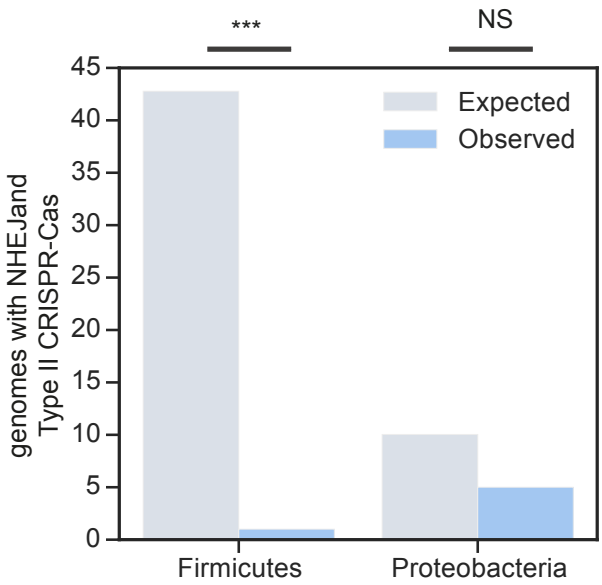
AGTTAAAAAGAATTTAAAGTCAAGAAGTA
ATGTTGATGGATCGTATCAAAGCGACATAC
AGGAATTGAGACACCTCAATATATACTTGC
ACACAAAGAAGTACATCAAGGGACAATTAC
CGAGCAAAGTTTCATCCGTTTAAATCAATA
TTAACGGTATGGAAGAAGCGAGTATCAATA
TACCGAATGAATTTTTAAAATATTCAGGCA
TCTTAAAGTTATTGAAGAAAGGTTATAACA
GGCAATGTTATTTTATCGGATTTTAAAAAC
GCTAATGACAGACCATTATTTGATGCTAAC
GACAAAATCGAACTATCATTAAAAGTTAAA
GCTATAGACGGAAGTTTCAACTTATTATAA
ACGACAGATATACGTCAGCGATTTATAATC
ACCGAATGAATTTTTAAAACATTCAGGCA
CAAATCTATTCAAGATACTATCGAAGCTGT



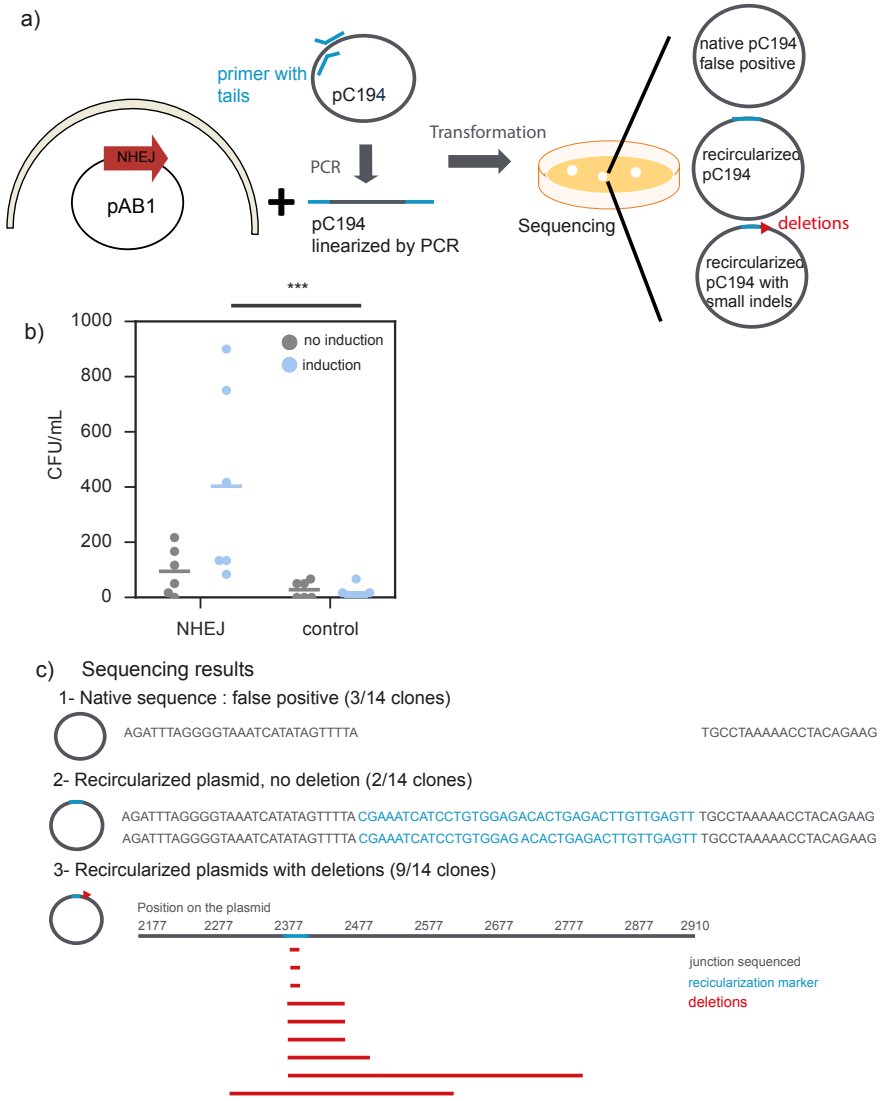
Supplementary Figure 1: Frequency of NHEJ and CRISPR-Cas systems by clades.



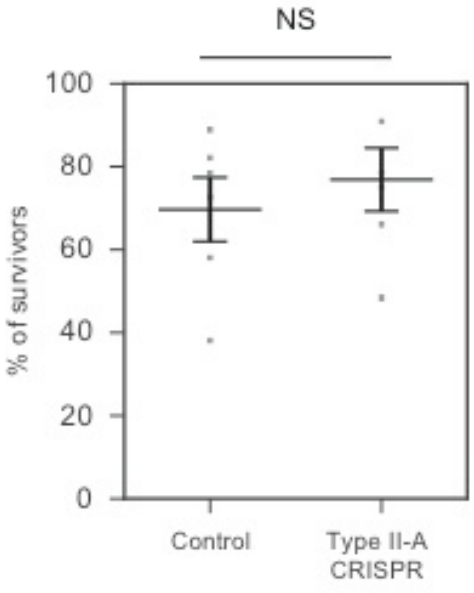
Supplementary Figure 2: Distribution of NHEJ and CRISPR-Cas systems in the genomes of Firmicutes and Proteobacteria. a, Frequency of NHEJ and CRISPR-Cas systems in the 1201 Firmicutes and 2988 Proteobacteria genomes. **b,** Frequency of NHEJ and CRISPR-Cas systems in function of genome size. The histogram on the bottom represents the distribution of genome sizes in each clade. The frequency represents the frequency of genomes carrying a system within the genome size range. Vertical line corresponds to the size cut-off (5Mb).



Supplementary Figure 3: Associations between NHEJ and Type II CRISPR-Cas systems in Firmicutes and Proteobacteria genomes. The expected values are the product of the marginal row and column totals divided by the grand total of the contingency tables. Statistics were calculated using Fisher's Exact Test. ** P< 0.05, *** P< 0.01

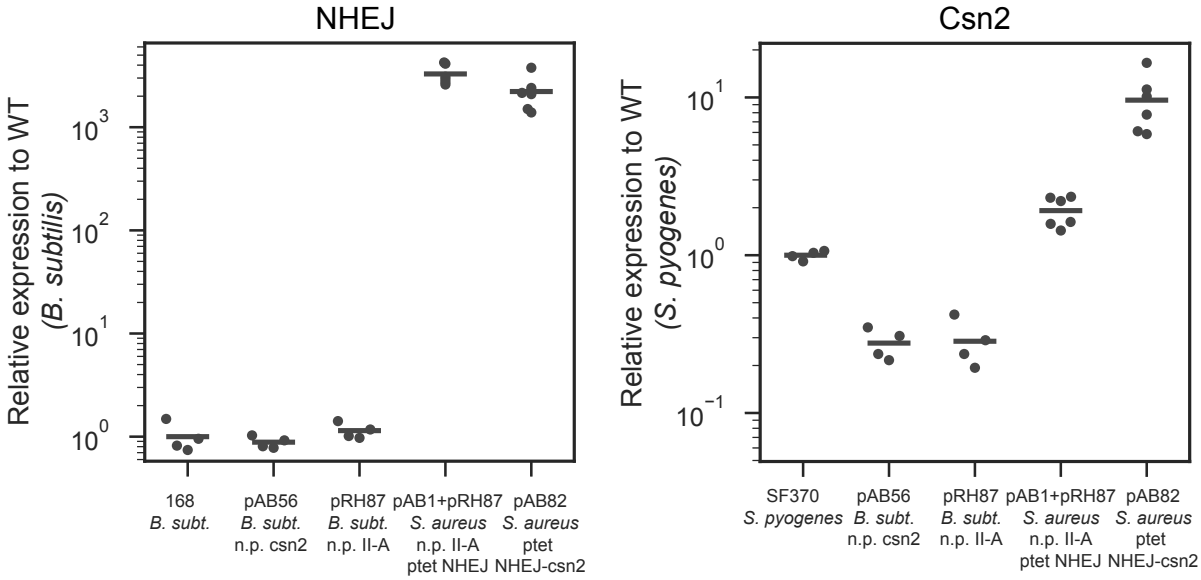


Supplementary Figure 4: NHEJ system of *B. subtilis* is functional in *S. aureus*. a, Introduction of a linearized plasmid providing resistance to chloramphenicol (pC194) into cells carrying the NHEJ system (pAB1) either in the presence or absence of aTc (induction) in *S. aureus*. Only cells that re-circularized the plasmid could form colonies on chloramphenicol plates. **b,** The CFU of five replicates is represented, error bars corresponds to s.e.m. *** P< 0.01 Wilcoxon test **c,** Sequencing results of transformants recovered after NHEJ induction. The primers used to linearize pC194 were designed to carry 5' tails (in blue). If plasmids were recircularized by the NHEJ system, we expected to find those tails sequence at the junction. Results are shown for clones recovered after NHEJ induction. Blue corresponds to overhangs and red to deletions. The insertion of the intact tail sequences at the junction was observed in 2/14 colonies recovered after NHEJ induction. In 9/14 colonies, small deletions were observed. The remaining 3/14 colonies were false positives. In contrast, all colonies obtained in the control experiment corresponded to the pC194 plasmid that was likely present in trace amounts in the PCR products used for the transformation.



Supplementary Figure 5 : Irradiation of *S. aureus* cells

S. aureus cells carrying a type II-A CRISPR-Cas systems or a control plasmid were irradiated. The survival rate was determined as the ration of CFUs obtained for irradiated over non-irradiated cells. No significant difference was observed (n=5 Wilcoxon, P=0.5).



Supplementary Figure 6 : Expression level of NHEJ and Csn2

Expression of NHEJ and Csn2 was measure using q-PCR in strains used in the study (see Supplementary Table 3 for a description of the plasmids). Expression was normalized to the 16s_rRNA expression in each strain measured using the same set of primers. Expression is shown relative to the wild-type: *B. subtilis* 168 for NHEJ and *S. pyogenes* SF370 for Csn2.

References

1. Horinouchi, S. & Weisblum, B. Nucleotide Sequence and Functional Map of pC194, a Plasmid That Specifies Inducible Chloramphenicol Resistance. *J. Bacteriol.* **150,** 815–825 (1982).
2. Byeon, W. H. & Weisblum, B. Post-transcriptional regulation of chloramphenicol acetyl transferase. *J. Bacteriol.* **158,** 543–550 (1984).
3. Khan, S. a, Adler, G. K. & Novick, R. P. Functional origin of replication of pT181 plasmid DNA is contained within a 168-base-pair segment. *Proc. Natl. Acad. Sci. U. S. A.* **79,** 4580–4584 (1982).
4. Ogasawara, N. Systematic function analysis of Bacillus subtilis genes. *Res. Microbiol.* **151,** 129–134 (2000).
5. de Vos, W. M. Gene cloning in lactic streptococci. *Netherlands Milk Dairy J.* **40,** 141–154 (1986).
6. Husmann, L. K., Scott, J. R., Lindahl, G. & Stenberg, L. Expression of the Arp protein, a member of the M protein family, is not sufficient to inhibit phagocytosis of Streptococcus pyogenes. *Infect. Immun.* **63,** 345–348 (1995).
7. Charpentier, E. *et al.* Novel Cassette-Based Shuttle Vector System for Gram-Positive Bacteria Novel Cassette-Based Shuttle Vector System for Gram-Positive Bacteria. *Appl. Environ. Microbiol.* **70,** 6076–6085 (2004).
8. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519,** 199–202 (2015).
9. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. a. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **4,** 633–637 (2014).

Box 4: Major points of Chapter 5

- **Type II-A** CRISPR-Cas systems and the **NHEJ** (Non Homologous End Joining) pathway **only co-occur once** among 5563 bacterial genomes.
- Exploration of the molecular mechanisms behind this negative association using **three different experimental models**: *Staphylococcus aureus*, *Streptococcus thermophilus* and *Bacillus subtilis*.
- **NHEJ** pathway has **no effect** type II-A CRISPR-Cas systems **interference or adaptation**
- **NHEJ repair is inhibited by the Csn2 protein** from type II-A CRISPR-Cas system.
- **First evidence of an antagonistic interaction** between a CRISPR-Cas system and a DNA repair pathway.

Potential **use of Csn2 to promote homologous recombination** in genome editing experiments in mammalian cells.

Chapter 6

Conclusions and perspectives

The goal of my thesis was to explore CRISPR-Cas interactions with DNA repair pathways to better understand CRISPR-Cas systems distribution in bacterial genomes. I first described precisely and quantitatively the distribution of CRISPR arrays and Cas clusters in bacterial genomes underlying the diversity and complexity of these immune systems in terms of taxonomic distribution and genetic organization. I thus provided evidence for a sparse distribution of CRISPR-Cas systems in bacterial genomes. Trying to explain this sparse distribution, I studied the co-occurrence patterns of CRISPR-Cas systems and DNA repair pathways. I observed many positive and negative associations between DNA repair and CRISPR-Cas systems leading me to propose a scenario for the impact of these associations on the distribution of CRISPR-Cas systems. Finally, by combining bioinformatics and experimental approaches, I introduced evidence for a negative interaction between a CRISPR-Cas system (Type II-A) and a DNA repair pathway (NHEJ). We suggest that the antagonism is caused by the inhibition of NHEJ repair by Csn2 a protein of the type II-A CRISPR-Cas systems. This confirmed that the observed co-occurrence pattern of avoidance was caused by a direct interaction.

From a methodology point of view, the CRISPR field has benefited a lot from mixed bioinformatical and experimental approaches. The first discoveries on CRISPR-Cas systems came from the studies of bacterial genomes [121, 217]. Today, new CRISPR-Cas variants like type V or type VI with important applications potential, were discovered through bioinformatics exploration of bacterial genomes followed by detailed molecular characterization [242, 312, 2]. In my PhD, the use of both approaches allowed me to study at the same time in a broad and specific manner my subject of interest. Understanding the underlying mechanisms of a specific interaction between a DNA repair pathway and a CRISPR-Cas system was essential to prove that the co-occurrence pattern we observed, was caused by a direct interaction and was not the consequence of another indirect association. Furthermore, bioinformatics allowed me to show that this specific case was not isolated but seemed more representative of diverse interactions between CRISPR-Cas systems and DNA repair pathways. The back and forth between both approaches was

also essential for deciding on research directions. In Chapter 5, the observation of a specific co-occurrence pattern led us to investigate experimentally its possible cause. Our first hypothesis was that the NHEJ repair pathway would repair DSB generated by Cas9. When first results showed that this was not the case, we kept looking for alternative hypotheses, only because we had the conviction that there was a direct interaction, thanks to the bioinformatics data.

A second important methodological aspect of this work was the use of multiple model bacteria in the experimental approaches namely, *Staphylococcus aureus*, *Bacillus subtilis* and *Streptococcus thermophilus*. We had observed that both systems dit not co-occur in bacterial genomes. Therefore, the study of the interaction between NHEJ and type II-A required to construct a *de facto* artificial experimental set up. In order to avoid artifacts, it became quickly apparent that results should be confirmed in several model bacteria. Most of the work done in the lab on type II CRISPR-Cas systems were performed on *S. aureus* as it is easy to modify genetically and many protocols to study CRISPR-Cas activity were developed for this model. The first experiments were therefore conducted with this model while being conscious of the limitations as *S. aureus* do not encode naturally neither NHEJ or a type II-A CRISPR-Cas system.

The use of this model actually misled us for some time as it led to an artifact caused by one of our expression systems. When using a specific expression system (the combination of the pE194 and a ptet promoter in *S. aureus*), the induction of NHEJ with aTc led to the abolishment of the ability to acquire new spacers. We had used the empty vector as a control. In these conditions, no impact on adaptation could be observed. This first let us conclude that the expression of NHEJ could in fact abolish adaptation. However, while performing complementary work, we realized that inducing the expression of any protein (like a GFP) using this expression system, leads to the observed phenotype of adaptation abolishment. We still do not understand the mechanism behind this intriguing result. However, this led us to change our expression system and also convinced us of the necessity to confirm our results in different model bacteria. A important challenge remained : the absence of existing protocols to assess the activity of both systems in bacteria naturally encoding one or the other system (*B. subtilis* for NHEJ, *S. thermophilus* for type II-A CRISPR-Cas systems.). We therefore decided to assess the activity of each system, in its natural context and mimick the arrival or the other system by horizontal gene transfer.

The study of CRISPR-Cas systems distribution revealed two new aspects of CRISPR-Cas systems diversity. While it was known that CRISPR-Cas systems are distributed unevenly in different environments and in different phyla [42], no precise analysis of the relative abundance of subtypes by clades had been performed. We provide a precise taxonomic distribution and show that subtypes of

CRISPR-Cas systems are distributed unevenly. We show that some subtypes are present in diverse clades while some others are clades specific. This suggests that some subtypes might accommodate more easily to a new genetic background than others or that certain subptypes developed some specialization.

A second key conclusion of the concomittant analysis of CRISPR arrays and Cas clusters is the importance of non-canonical forms of CRISPR-Cas systems. As mentioned in the introduction, CRISPR-Cas systems are generally defined as the association of one CRISPR array with a cluster of *cas* genes. Orphan or distant elements [19, 272] had been described as well as loci involving several Cas clusters and CRISPR arrays [256, 161] ; however, no quantitative measurements of these different elements had been performed. I showed that while many orphan and distant CRISPR arrays can be detected in bacterial genomes, orphan and alone Cas clusters are very rare. Moreover, those distant and orphan arrays encode fewer repeats than their counterparts which are part of CRISPR-Cas loci suggesting a functional difference between those three types of elements. These orphan and distant arrays could be remains of previously active CRISPR-Cas systems that could be reactivated upon a transfer of a new set of *cas* genes. As CRISPR-Cas systems have been shown to be involved in other bacterial processes beyond immunity [296], orphan arrays could also serve other functions.

While orphan elements represent one side of the spectrum of CRISPR-Cas organization, the other extreme, the very complex CRISPR-Cas loci is no less intriguing. The advantage of encoding complex CRISPR-Cas loci have yet to be fully explored. Recently, it was argued that encoding several CRISPR arrays could be adaptive as it would allow a trade off between memory span and learning speed [290]. Several studies have recently started to explore why encoding several Cas clusters could also be advantageous [189, 136, 248].

First, some type III CRISPR-Cas systems harbours Csm6 which function remained until recently unknown. Csm6 is a non specific RNAse which is activated by a small molecule : cyclic oligoadenylates or coA [189, 136]. CoA is produced by Cas10, a protein of the effector complex of type III systems, in presence of a target. The accumulation of cOA constitutes an intracellular signal that infection has not been prevented. It was therefore hypothesized that CRISPR-Cas immunity could function in several phases and involve several types of CRISPR-Cas systems [5]. In bacteria harboring type I and type III systems, upon a phage infection, the phage would be first targeted by the type I system. If however the phage would be immune to the type I system through for example a mutation in the PAM ; the phage could then be targeted by the type III system which does not rely on PAM recognition. This would constitute a second line of defense. Once activated, Cas10 also produces cOA. If the second line of defense fails, Cas10 will keep on producing cOA which will in the end lead to cell death or dormancy thus forming a third line

of defense to prevent further infection of the population [5]. This scenario implies the association of type I and III systems.

In our analysis, we observed patterns of co-occurrence pointing to preferred associations between type I and type III systems. Furthermore, our analysis of large clusters of Cas proteins implicates only proteins of type I and III systems. Such a physical association constitutes another clue of the potential functional linkage between those types of CRISPR-Cas systems. Our analysis also reveals preferred associations between subtypes such as type I-F and III-B. Recently, it was demonstrated that in *Marinomonas mediterranea* MMB-1, type III-B systems can process crRNA from another type I-F system present in the cell when type I-F is impaired [247]. This could constitute another advantage of encoding several Cas clusters and complex CRISPR-Cas loci. Those two experimental examples and our data which demonstrates that these types of associations are not rare reveal how CRISPR-Cas systems generate complex defense strategies. More studies on experimental models encoding several Cas clusters and arrays could help reveal new complex defense approaches of bacteria.

The ability of type III-B to process spacers from the type I-F also questions the specificity of subtypes to diverse repeats. The absence of degenerated repeats in CRISPR arrays would suggest a limited possibility for a specific subtype to deal with diverse repeats. The subtype prediction we performed using only CRISPR repeats also underlines some specificity between repeats and subtypes. More precise studies of repeats of type I-F and type III-B systems as well as experimental studies on the permissiveness of the type III-B to different repeats could help understand the basic requirements of Cas proteins to efficiently process crRNAs.

Predicting subtypes from repeats could also provide a usefool tool for the analysis of CRISPR-Cas systems in metagenomes. One main constraint when studying CRISPR-Cas in metagenomes is the absence of contigs with complete Cas clusters thus strongly limiting any analysis at the subtype level. Several metagenomics analysis have tried to circumvent this issue by for example using inly Cas1 to assign subtypes. [43, 42]. However, some subtypes do not encode Cas1 and this methodology still requires the detection of *cas* genes. Many hypothesis have been made on the prevalence of CRISPR-Cas systems in different environments [294] such as an enrichment in thermophilic environments. Allowing subtype assignment on repeats only, paves the way for subtype quantification in different environments. This would probably bring new clues to the ecological factors that could favor one subtype over the others.

My work focused on CRISPR-Cas systems interactions with DNA repair pathway. The most well known interaction between a DNA repair pathway and a CRISPR-Cas system came through the study of the role of RecBCD in producing

prespacers for type I-E systems [149]. A subsequent study showed that other DNA repair proteins could help type I-E systems to acquire new systems [122]. I provide the first example of a negative interaction between a CRISPR-Cas system and a DNA repair pathway. My results suggest that when a type II-A CRISPR-Cas system is transfered to a bacteria encoding NHEJ, Csn2 will inhibit NHEJ repair which will in the end lead to the CRISPR-Cas loss. Those two experimentally confirmed associations represent extremes of the spectrum of potential interactions between DNA repair pathways and CRISPR-Cas systems with a dependency on one hand (Type I-E RecBCD) and an incompatibility on the other. Further interactions which could be experimentally characterized could fall in either categories or in between.

For both interactions (type I-E RecBCD and type II-A NHEJ), subtype specificity is essential as the interaction does not hold true for other related subtypes. This subtype specificity seems to be common to other associations. When studying other patterns of co-occurrence, it is striking to note that very different subtypes share similar associations with DNA repair pathways (like type II-C, I-B and III-D) while systems belonging to the same types like type II or some type I do not. This underlines how those interactions could cause the sparse distribution observed at the subtype level.

Such interactions between CRISPR-Cas systems and chromosomal encoded systems could concern other functions than DNA repair. One particularly interesting set of functions and potential associations to study would be with other defense systems. It was observed that genomes encoding R-M systems are more likely to encode CRISPR-Cas systems [196]. In *S. thermophilus*, type II CRISPR-Cas systems and type II R-M systems systems work synergistically to prevent infection by phages [68]. Studying the compatibility of both systems at a fine level could help reveal cooperation or incompatibilities between defense systems and bring elements to answer the question "Is more immunity best?".

When working in the CRISPR field, new findings are systematically questioned through the point of view of potential applications. Even if my main research question remains very fundamental, some conclusions could help improve existing biotechnologies. CRISPR-Cas systems are mainly used in microbes in three manners : as genome editing technologies, antimicrobials or for gene regulation purposes [185, 18, 29]. Interactions between DNA repair pathways and CRISPR-Cas systems are at the heart of the two first applications. For genome editing purposes, the goal is to generate breaks using CRISPR-Cas systems that would then be repaired by a DNA repair pathways either endogenous to the host or exogenous and brought with the CRISPR-Cas system. On the contrary, for antimicrobials, the objective is to generate cuts that won't be repaired and will therefore lead to cell death.

Many genome editing techniques in bacteria rely on the introduction of an exogenous NHEJ pathway. Given my results, when working with a new organisms, one should make sure that it does not encode Csn2. On the contrary, in eukaryotes, one current issue is that because NHEJ is always present making the introduction of mutations through homologous recombination difficult. In that context, introducing Csn2 might help promote homologous recombination. However, it should be first demonstrated that Csn2 also inhibits eukaryotic NHEJ. Experiment to confirm this inhibition is currently undertaken in collaboration with another lab.

I also observed other negative associations of NHEJ with specific CRISPR-Cas systems like with the type I-E. This negative association should be investigated experimentally to understand the underlying mechanism. As type I-E is the most common subtype in bacteria, this could help improve genome editing strategies for bacteria encoding this CRISPR-Cas system. Similarly, those interactions should be taken into account when trying to generate specific antimicrobials. Playing with the incompatibilities of specific DNA repair pathways and CRISPR-Cas systems might help improve their efficiencies. More specifically introducing Csn2 with Cas9 when bacteria encode NHEJ might help improve killing efficiency.

Finally, I want to go back to the question that started my PhD : Why are CRISPR-Cas systems relatively rare given their role as an immune system and the fact that they are frequently transfered horizontally ? As already explained, several hypotheses had been put forward to explain this relative scarcity of CRISPR-Cas systems. First, autoimmunity: the acquisition of a self-targeting spacer leads in the vast majority of cases to cell death [110]. However, given the diverse taxonomic distribution detailed in Chapter 3, there is no clear reason why the cost of autoimmunity should vary between clades. Second, harboring general defenses like restriction modification or surface modification can be more advantageous than encoding a specialized defense system like CRISPR-Cas [296]. However, this would also apply to archaea and therefore fail to explain why 90% of archaea encode CRISPR-Cas systems and only 40% of bacteria. Third, by limiting horizontal gene transfer (HGT), CRISPR-Cas systems can prevent the uptake of advantageous mobile genetic elements [128]. This is true for other defense systems like restriction modification of which two copies are present on average per bacterial genome [196]. I introduced a novel and complementary hypothesis : the possibility that the success of CRISPR-Cas acquisition by horizontal gene transfer is partly determined by the interactions of these systems with the genetic background of the host. Using interactions with DNA repair pathways, I demonstrated that those interactions impact CRISPR-Cas systems distribution in bacterial genomes. However, the importance of the genetic background is not limited to DNA repair pathways. Other elements could influence the distribution of CRISPR-Cas systems. I provide a method to detect significant associations, which could be adapted to the study of other elements. For example, as mentioned above, it would be interesting to

focus on the interplay with other defense systems.

To conclude, I hope this work underlines the diversity and complexity of CRISPR-Cas systems and bring new elements to the understanding of their evolution.

Bibliography

[1] Sophie S Abby, Bertrand Néron, Hervé Ménager, Marie Touchon, and Eduardo P C Rocha, *MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems.*, PloS one **9** (2014), no. 10, e110726.

[2] Omar O Abudayyeh, Jonathan S Gootenberg, Silvana Konermann, Julia Joung, Ian M Slaymaker, David BT Cox, Sergey Shmakov, Kira S Makarova, Ekaterina Semenova, Leonid Minakhin, Konstantin Severinov, Aviv Regev, Eric S Lander, Eugene V Koonin, and Feng Zhang, *C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector*, Science **353** (2016), no. 6299, aaf5573.

[3] Yoshihiro Agari, Keiko Sakamoto, Masatada Tamakoshi, Tairo Oshima, Seiki Kuramitsu, and Akeo Shinkai, *Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection*, Journal of Molecular Biology **395** (2010), no. 2, 270–281.

[4] Juan C Alonso, Paula P Cardenas, Humberto Sanchez, James Hejna, Yuki Suzuki, and Kunio Takeyasu, *Early steps of double-strand break repair in Bacillus subtilis*, DNA Repair **12** (2013), no. 3, 162–176.

[5] Gil Amitai and Rotem Sorek, *Intracellular signaling in CRISPR-Cas defense*, Science **357** (2017), no. 6351, 550–551.

[6] Susan K. Amundsen, Jutta Fero, Lori M. Hansen, Gareth A. Cromie, Jay V. Solnick, Gerald R. Smith, and Nina R. Salama, *Helicobacter pylori AddAB helicase-nuclease and RecA promote recombination-related DNA repair and survival during stomach colonization*, Molecular Microbiology **69** (2008), no. 4, 994–1007.

[7] Carolin Anders, Ole Niewoehner, Alessia Duerst, and Martin Jinek, *Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease*, Nature **513** (2014), no. 7519, 569–573.

[8] Rika E. Anderson, William J. Brazelton, and John A. Baross, *Using CRISPRs as ametagenomic tool to identify microbial hosts of a diffuse flow*

hydrothermal vent viral assemblage, FEMS Microbiology Ecology **77** (2011), no. 1, 120–133.

[9] Anders F. Andersson and Jillian F. Banfield, *Virus population dynamics and acquired virus resistance in natural microbial communities*, Science **320** (2008), no. 5879, 1047–1050.

[10] Jideofor Aniukwu, Michael S. Glickman, and Stewart Shuman, *The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends*, Genes and Development **22** (2008), no. 4, 512–527.

[11] S Anupama, Aswathy Rajan Mp, M Gurusaran, P Radha, Dinesh Kumar Ks, R Chitra, Hima Vyshanavi Am, and K Sekar, *Evolutionary Analysis of CRISPRs in Archaea : An Evidence for Horizontal*, Journal of Proteomics & Bioinformatics **S9** (2014), no. 005.

[12] L. Aravind and E. V. Koonin, *Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system*, Genome Research **11** (2001), no. 8, 1365–1374.

[13] Zihni Arslan, Veronica Hermanns, Reinhild Wurm, Rolf Wagner, and Ümit Pul, *Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system.*, Nucleic Acids Research **42** (2014), no. 12, 7884–93.

[14] Zihni Arslan, Reinhild Wurm, Oleksandr Brener, Philipp Ellinger, Luitgard Nagel-Steger, Filipp Oesterhelt, Lutz Schmitt, Dieter Willbold, Rolf Wagner, Holger Gohlke, Sander H J Smits, and Ümit Pul, *Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2*, Nucleic Acids Research **41** (2013), no. 12, 6347–6359.

[15] Silvia Ayora, Begoña Carrasco, Paula P. Cárdenas, Carolina E. César, Cristina Cañas, Tribhuwan Yadav, Chiara Marchisone, and Juan C. Alonso, *Double-strand break repair in bacteria: A view from Bacillus subtilis*, FEMS Microbiology Reviews **35** (2011), no. 6, 1055–1081.

[16] Mohan Babu, Natalia Beloglazova, Robert Flick, Chris Graham, Tatiana Skarina, Boguslaw Nocek, Alla Gagarinova, Oxana Pogoutse, Greg Brown, Andrew Binkowski, Sadhna Phanse, Andrzej Joachimiak, Eugene V Koonin, Alexei Savchenko, Andrew Emili, Jack Greenblatt, Aled M Edwards, and Alexander F Yakunin, *A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair.*, Molecular microbiology **79** (2011), no. 2, 484–502.

[17] Rodolphe Barrangou, *The roles of CRISPR Cas systems in adaptive immunity and beyond*, Current Opinion in Immunology **32** (2015), 36–41.

[18] Rodolphe Barrangou and Jennifer A Doudna, *Applications of CRISPR technologies in research and beyond*, Nature Biotechnology **34** (2016), no. 9, 933–941.

[19] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis a Romero, and Philippe Horvath, *CRISPR provides acquired resistance against viruses in prokaryotes.*, Science **315** (2007), no. 5819, 1709–12.

[20] Rodolphe Barrangou and Philippe Horvath, *A decade of discovery: CRISPR functions and applications*, Nature Microbiology **2** (2017), no. June, 17092.

[21] Maria Paloma S Barros, Camila T França, Rosanny Holanda F B Lins, Milena Danda V Santos, Ednaldo J Silva, Maria Betânia M Oliveira, Vladimir M Silveira-Filho, Antônio M Rezende, Valdir Q Balbino, and Tereza Cristina Leal-Balbino, *Dynamics of CRISPR Loci in Microevolutionary Process of Yersinia pestis Strains.*, PloS one **9** (2014), no. 9, e108353.

[22] Pierre Beguin, Nicole Charpin, Eugene V. Koonin, Patrick Forterre, and Mart Krupovic, *Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems*, Nucleic Acids Research **44** (2016), no. 21, 10367–10376.

[23] Alex Van Belkum, Leah B Soriaga, Matthew C Lafave, Srividya Akella, Jean-baptiste Veyrieras, E Magda Barbu, Dee Shortridge, Bernadette Blanc, Gregory Hannum, Gilles Zambardi, Kristofer Miller, Mark C Enright, Nathalie Mugnier, Daniel Brami, Stéphane Schicklin, Martina Felderman, Ariel S Schwartz, Toby H Richardson, Todd C Peterson, Bolyn Hubby, and Kyle C Cady, *Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic- Resistant Pseudomonas aeruginosa*, mBio **6** (2015), no. 6, 1–13.

[24] Christian Benda, Judith Ebert, Richard A. Scheltema, Herbert B. Schiller, Marc Baumgärtner, Fabien Bonneau, Matthias Mann, and Elena Conti, *Structural Model of a CRISPR RNA-Silencing Complex Reveals the RNA-Target Cleavage Activity in Cmr4*, Molecular Cell **56** (2014), no. 1, 43–54.

[25] Faina S Berezovskaya, Yuri I Wolf, Eugene V Koonin, and Georgy P Karev, *Pseudo-chaotic oscillations in CRISPR-virus coevolution predicted by bifurcation analysis.*, Biology direct **9** (2014), no. 1, 13.

[26] Margret E. Berg Miller, Carl J. Yeoman, Nicholas Chia, Susannah G. Tringe, Florent E. Angly, Robert A. Edwards, Harry J. Flint, Raphael Lamed, Edward A. Bayer, and Bryan A. White, *Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome*, Environmental Microbiology **14** (2012), no. 1, 207–227.

[27] Aude Bernheim, Alicia Calvo Villamanan, Clovis Basier, David Bikard, Microbial Evolutionary Genomics, and Synthetic Biology Group, *Inhibition of NHEJ repair by type II-A CRISPR-Cas systems*, bioRxiv (2017).

[28] P R Bianco and S C Kowalczykowski, *The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5'-GCTGGTGG-3'.*, Proceedings of the National Academy of Sciences of the United States of America **94** (1997), no. 13, 6706–11.

[29] David Bikard and Rodolphe Barrangou, *Using CRISPR-Cas systems as antimicrobials*, Current Opinion in Microbiology **37** (2017), 155–160.

[30] David Bikard, Chad W Euler, Wenyan Jiang, Philip M Nussenzweig, Gregory W Goldberg, Xavier Duportet, Vincent a Fischetti, and Luciano a Marraffini, *Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials.*, Nature biotechnology **32** (2014), no. 11, 1146–1150.

[31] David Bikard, Asma Hatoum-Aslan, Daniel Mucida, and Luciano a Marraffini, *CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection.*, Cell host & microbe **12** (2012), no. 2, 177–86.

[32] Charles Bland, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, and Philip Hugenholtz, *CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats*, BMC Bioinformatics **8** (2007), no. 1, 209.

[33] Timothy R. Blosser, Luuk Loeff, Edze R. Westra, Marnix Vlot, Tim Künne, Małgorzata Sobota, Cees Dekker, Stan J.J. Brouns, and Chirlmin Joo, *Two Distinct DNA Binding Modes Guide Dual Roles of a CRISPR-Cas Protein Complex*, Molecular Cell **58** (2015), no. 1, 60–70.

[34] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich, *Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin*, Microbiology **151** (2005), no. 8, 2551–2561.

[35] Joe Bondy-Denomy, April Pawluk, Karen L Maxwell, and Alan R Davidson, *Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system.*, Nature **493** (2013), no. 7432, 429–32.

[36] Joseph Bondy-denomy, Karen L Maxwell, Bianca Garcia, and Maryclare Rollins, *Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins*, Nature **526** (2015), no. 7571, 136–139.

[37] Adair L Borges, Alan R Davidson, and Joseph Bondy-denomy, *The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs*, Annual Review of Virology **4** (2017).

[38] Richard Bowater and Aidan J Doherty, *Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining.*, PLoS genetics **2** (2006), no. 2, e8.

[39] Alexandra E Briner and Rodolphe Barrangou, *Deciphering and shaping bacterial diversity through CRISPR*, Current Opinion in Microbiology **31** (2016), 101–108.

[40] Stan J J Brouns, Matthijs M Jore, Magnus Lundgren, Edze R Westra, Rik J H Slijkhuis, Ambrosius P L Snijders, Mark J Dickman, Kira S Makarova, Eugene V Koonin, and John van der Oost, *Small CRISPR RNAs guide antiviral defense in prokaryotes.*, Science **321** (2008), no. 5891, 960–4.

[41] Alexandra Bryson, Young Hwang, Scott Sherrill-Mix, Gary D. Wu, James D. Lewis, Lindsay Black, Tyson A. Clark, and Frederic D. Bushman, *Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR- Cas9*, mBio **6** (2015), no. 32, 1–9.

[42] D Burstein, LC Sun, CT Brown, I Sharon, K Anantharaman, AJ Probst, BC Thomas, and JB Banfield, *Major bacterial lineages are essentially devoid of CRISPR-Cas viral defense systems.*, Nature Communications **7** (2016), 10613.

[43] David Burstein, Lucas B Harrington, Steven C Strutt, and Alexander J Probst, *New CRISPR-Cas systems from uncultivated microbes*, Nature **542** (2016), no. 7640, 237–241.

[44] Kyle C. Cady and George A. OToole, *Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins*, Journal of Bacteriology **193** (2011), no. 14, 3433–3445.

[45] Jason Carte, Ruiying Wang, Hong Li, Rebecca M Terns, and Michael P Terns, *Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.*, Genes & development **22** (2008), no. 24, 3489–96.

[46] Guoshi Chai, Min Yu, Lixu Jiang, Yaocong Duan, and Jian Huang, *HMM-CAS: a web tool for the identification and domain annotations of Cas proteins*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **5963** (2017).

[47] Sajib Chakraborty, Ambrosius P. Snijders, Rajib Chakravorty, Musaddeque Ahmed, Ashek Md Tarek, and M. Anwar Hossain, *Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the*

horizontal transfer of CRISPR locus among bacteria, Molecular Phylogenetics and Evolution **56** (2010), no. 3, 878–887.

[48] Romain Chayot, Benjamin Montagne, Didier Mazel, and Miria Ricchetti, *An end-joining repair mechanism in Escherichia coli.*, Proceedings of the National Academy of Sciences of the United States of America **107** (2010), no. 5, 2141–6.

[49] Lauren M Childs, Whitney E England, Mark J Young, Joshua S Weitz, and Rachel J Whitaker, *CRISPR-Induced Distributed Immunity in Microbial Populations.*, PloS one **9** (2014), no. 7, e101710.

[50] Lauren M. Childs, Nicole L. Held, Mark J. Young, Rachel J. Whitaker, and Joshua S. Weitz, *Multiscale model of crispr-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin*, Evolution **66** (2012), no. 7, 2015–2029.

[51] Saikat Chowdhury, Joshua Carter, Mary Clare F. Rollins, Sarah M. Golden, Ryan N. Jackson, Connor Hoffmann, Lyn'Al Nosaka, Joseph Bondy-Denomy, Karen L. Maxwell, Alan R. Davidson, Elizabeth R. Fischer, Gabriel C. Lander, and Blake Wiedenheft, *Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex*, Cell **169** (2017), no. 1, 47–57.

[52] Robert J Citorik, Mark Mimee, and Timothy K Lu, *Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases*, Nature Biotechnology **32** (2014), no. 11, 1141–1145.

[53] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano a Marraffini, and Feng Zhang, *Multiplex genome engineering using CRISPR/Cas systems.*, Science **339** (2013), no. 6121, 819–23.

[54] Michael M Cox, *Motoring along with the bacterial RecA protein.*, Nature Reviews Molecular Cell Biology **8** (2007), no. 2, 127–138.

[55] ———, *Regulation of Bacterial RecA Protein Function*, Critical Reviews in Biochemistry and Molecular Biology **42** (2007), 41–63.

[56] Gareth A. Cromie, *Phylogenetic ubiquity and shuffling of the bacterial RecBCD and AddAB recombination complexes.*, Journal of bacteriology **191** (2009), no. 16, 5076–84.

[57] Lun Cui and David Bikard, *Consequences of Cas9 cleavage in the chromosome of Escherichia coli*, Nucleic Acids Research **44** (2016), no. 9, 4243–4251.

[58] Kirill a Datsenko, Ksenia Pougach, Anton Tikhonov, Barry L Wanner, Konstantin Severinov, and Ekaterina Semenova, *Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system.*, Nature communications **3** (2012), no. May, 945.

[59] Nigel F. Delaney, Susan Balenger, Camille Bonneaud, Christopher J. Marx, Geoffrey E. Hill, Naola Ferguson-Noel, Peter Tsai, Allen Rodrigo, and Scott V. Edwards, *Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, Mycoplasma Gallisepticum*, PLoS Genetics **8** (2012), no. 2, e1002511.

[60] Marina Della, Phillip L Palmbos, Hui-Min Tseng, Louise M Tonkin, James M Daley, Leana M Topper, Robert S Pitcher, Alan E Tomkinson, Thomas E Wilson, and Aidan J Doherty, *Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine.*, Science **306** (2004), no. 5696, 683–685.

[61] 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 **471** (2011), no. 7340, 602–607.

[62] Hélène Deveau, Rodolphe Barrangou, Josiane E. Garneau, Jessica Labonté, Christophe Fremaux, Patrick Boyaval, Dennis A. Romero, Philippe Horvath, and Sylvain Moineau, *Phage response to CRISPR-encoded resistance in Streptococcus thermophilus*, Journal of Bacteriology **190** (2008), no. 4, 1390–1400.

[63] César Díez-Villaseñor, Noemí M Guzmán, Cristóbal Almendros, Jesús García-Martínez, and Francisco J M Mojica, *CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli.*, RNA biology **10** (2013), no. 5, 792–802.

[64] Mark S Dillingham and Stephen C Kowalczykowski, *RecBCD enzyme and the repair of double-stranded DNA breaks.*, Microbiology and molecular biology reviews : MMBR **72** (2008), no. 4, 642–71.

[65] De Dong, Minghui Guo, Sihan Wang, Yuwei Zhu, Shuo Wang, Zhi Xiong, Jianzheng Yang, Zengliang Xu, and Zhiwei Huang, *Structural basis of CRISPR SpyCas9 inhibition by an anti-CRISPR protein*, Nature **546** (2017), no. 7658, 436–439.

[66] Jennifer Doudna and Samuel H Sternberg, *A Crack in Creation: Gene Editing and the Unthinkable Power to Control Evolution*, 2017.

[67] J W Drake, *A constant rate of spontaneous mutation in DNA-based microbes.*, Proceedings of the National Academy of Sciences of the United States of America **88** (1991), no. 16, 7160–4.

[68] Marie-Ève Dupuis, Manuela Villion, Alfonso H Magadán, and Sylvain Moineau, *CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance.*, Nature communications **4** (2013), no. May, 2087.

[69] Alexandra East-Seletsky, Mitchell R. O'Connell, Spencer C. Knight, David Burstein, Jamie H. D. Cate, Robert Tjian, and Jennifer A. Doudna, *Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection*, Nature **538** (2016), no. 7624, 270–273.

[70] Sean R. Eddy, *Profile hidden Markov models.*, Bioinformatics **14** (1998), no. 9, 755–763.

[71] Robert C Edgar, *PILER-CR: fast and accurate identification of CRISPR repeats.*, BMC bioinformatics **8** (2007), 18.

[72] Rotem Edgar and Udi Qimron, *The Escherichia coli CRISPR system protects from λ lysogenization, lysogens, and prophage induction*, Journal of Bacteriology **192** (2010), no. 23, 6291–6294.

[73] Jonathan A. Eisen and Philip C. Hanawalt, *A phylogenomic study of DNA repair genes, proteins, and processes.*, vol. 435, 1999.

[74] Philipp Ellinger, Zihni Arslan, Reinhild Wurm, Britta Tschapek, Colin MacKenzie, Klaus Pfeffer, Santosh Panjikar, Rolf Wagner, Lutz Schmitt, Holger Gohlke, Ümit Pul, and Sander H J Smits, *The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae*, Journal of Structural Biology **178** (2012), no. 3, 350–362.

[75] Joanne B. Emerson, Karen Andrade, Brian C. Thomas, Anders Norman, Eric E. Allen, Karla B. Heidelberg, and Jillian F. Banfield, *Virus-host and CRISPR dynamics in archaea-dominated hypersaline Lake tyrrell, Victoria, Australia*, Archaea **370871** (2013).

[76] Whitney E England and Rachel J Whitaker, *Evolutionary causes and consequences of diversified CRISPR immune profiles in natural populations.*, Biochemical Society transactions **41** (2013), no. 6, 1431–6.

[77] Susanne Erdmann and Roger A. Garrett, *Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms*, Molecular Microbiology **85** (2012), no. 6, 1044–1056.

[78] Michael A. Estrella, Fang Ting Kuo, and Scott Bailey, *RNA-activated DNA cleavage by the Type III-B CRISPR Cas effector complex*, Genes and Development **30** (2016), no. 4, 460–470.

[79] Joseph Felsenstein, *Phylogenies and the Comparative Method*, The American Naturalist **125** (1985), no. 1, 1–15.

[80] Alan Filipski, Oscar Murillo, Anna Freydenzon, Koichiro Tamura, and Sudhir Kumar, *Prospects for building large timetrees using molecular data with incomplete gene coverage among species*, Molecular Biology and Evolution **31** (2014), no. 9, 2542–2550.

[81] Peter C Fineran, Matthias J H Gerritzen, María Suárez-Diez, Tim Künne, Jos Boekhorst, Sacha A F T van Hijum, Raymond H J Staals, and Stan J J Brouns, *Degenerate target sites mediate rapid primed CRISPR adaptation.*, Proceedings of the National Academy of Sciences of the United States of America **111** (2014), no. 16, E1629–38.

[82] Ines Fonfara, Hagen Richter, Majda Bratovič, Anaïs Le Rhun, and Emmanuelle Charpentier, *The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA.*, Nature (2016), 1–19.

[83] Kenji Fukui, Noriko Nakagawa, Yoshiaki Kitamura, Yuya Nishida, Ryoji Masui, and Seiki Kuramitsu, *Crystal structure of Muts2 endonuclease domain and the mechanism of homologous recombination suppression*, Journal of Biological Chemistry **283** (2008), no. 48, 33417–33427.

[84] Salvatore Fusco, Rossana Liguori, Danila Limauro, Simonetta Bartolucci, Qunxin She, and Patrizia Contursi, *Transcriptome analysis of Sulfolobus solfataricus infected with two related fuselloviruses reveals novel insights into the regulation of CRISPR-Cas system*, Biochimie **118** (2015), 322–332.

[85] Roberto Galletto and Stephen C Kowalczykowski, *RecA*, Current Biology **17** (2007), no. 11, 395–397.

[86] Enriqueta García-Gutiérrez, Cristóbal Almendros, Francisco J. M. Mojica, Noemí M. Guzmán, and Jesús García-Martínez, *CRISPR Content Correlates with the Pathogenic Potential of Escherichia coli*, Plos One **10** (2015), no. 7, e0131935.

[87] Josiane E Garneau, Marie-Ève Dupuis, Manuela Villion, Dennis a Romero, Rodolphe Barrangou, Patrick Boyaval, Christophe Fremaux, Philippe Horvath, Alfonso H Magadán, and Sylvain Moineau, *The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA.*, Nature **468** (2010), no. 7320, 67–71.

[88] Giedrius Gasiunas, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys, *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria.*, Proceedings of the National Academy of Sciences of the United States of America **109** (2012), no. 39, E2579–86.

[89] Ruiquan Ge, Guoqin Mai, Pu Wang, Manli Zhou, Youxi Luo, Yunpeng Cai, and Fengfeng Zhou, *CRISPRdigger: detecting CRISPRs with better direct repeat annotations*, Scientific Reports **6** (2016), 32942.

[90] James S Godde and Amanda Bickerton, *The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.*, Journal of molecular evolution **62** (2006), no. 6, 718–29.

[91] Anna A Gogleva, Mikhail S Gelfand, and Irena I Artamonova, *Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs.*, BMC genomics **15** (2014), no. 1, 202.

[92] Gregory W Goldberg, Wenyan Jiang, David Bikard, and Luciano a Marraffini, *Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting.*, Nature **514** (2014), no. 7524, 633–637.

[93] Chunling Gong, Paola Bongiorno, Alexandra Martins, Nicolas C Stephanou, Hui Zhu, Stewart Shuman, and Michael S Glickman, *Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C*, Nature Structural & Molecular Biology **12** (2005), no. 4, 304–312.

[94] Uri Gophna, David M Kristensen, Yuri I Wolf, Ovidiu Popa, Christine Drevet, and Eugene V Koonin, *No evidence of inhibition of horizontal gene transfer by CRISPR Cas on evolutionary timescales*, The ISME Journal **9** (2015), no. 9, 2021–2027.

[95] Moran G. Goren, Shany Doron, Rea Globus, Gil Amitai, Rotem Sorek, and Udi Qimron, *Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array*, Cell Reports **16** (2016), no. 11, 2811–2818.

[96] Ibtissem Grissa, Gilles Vergnaud, and Christine Pourcel, *CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.*, Nucleic acids research **35** (2007), no. Web Server issue, W52–7.

[97] Felizza F. Gunderson and Nicholas P. Cianciotto, *The CRISPR-associated gene cas2 of Legionella pneumophila is required for intracellular infection of amoebae*, mBio **4** (2013), no. 2, e00074–13.

[98] Felizza F. Gunderson, Celeste A. Mallama, Stephanie G. Fairbairn, and Nicholas P. Cianciotto, *Nuclease activity of Legionella pneumophila Cas2*

promotes intracellular infection of amoebal host cells, Infection and Immunity **83** (2015), no. 3, 1008–1018.

[99] Peng Guo, Qiuxiang Cheng, Pengfei Xie, Yun Fan, Weihong Jiang, and Zhongjun Qin, *Characterization of the multiple CRISPR loci on Streptomyces linear plasmid pSHK1*, Acta Biochimica et Biophysica Sinica **43** (2011), no. 8, 630–639.

[100] Richa Gupta, Daniel Barkan, Gil Redelman-Sidi, Stewart Shuman, and Michael S. Glickman, *Mycobacteria exploit three genetically distinct DNA double-strand break repair pathways*, Molecular Microbiology **79** (2011), no. 2, 316–330.

[101] Jan O. Haerter and Kim Sneppen, *Spatial structure and Lamarckian adaptation explain extreme genetic diversity at CRISPR locus.*, mBio **3** (2012), no. 4, 1–6.

[102] Jan O. Haerter, Ala Trusina, and Kim Sneppen, *Targeted Bacterial Immunity Buffers Phage Diversity*, Journal of Virology **85** (2011), no. 20, 10554–10560.

[103] Asma Hatoum-Aslan and Luciano Marraffini, *Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens.*, Current opinion in microbiology **17** (2014), 82–90.

[104] Robert P. Hayes, Yibei Xiao, Fran Ding, Paul B. G. van Erp, Kanagalaghatta Rajashankar, Scott Bailey, Blake Wiedenheft, and Ailong Ke, *Structural basis for promiscuous PAM recognition in type IE Cascade from E. coli*, Nature **530** (2016), no. 7591, 499–503.

[105] Jiankui He and Michael W. Deem, *Heterogeneous diversity of spacers within CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)*, Physical Review Letters **105** (2010), no. 12, 128102.

[106] John F. Heidelberg, William C. Nelson, Thomas Schoenfeld, and Devaki Bhaya, *Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes*, PLoS ONE **4** (2009), no. 1, e4169.

[107] Nicole L. Held and Rachel J. Whitaker, *Viral biogeography revealed by signatures in Sulfolobus islandicus genomes*, Environmental Microbiology **11** (2009), no. 2, 457–466.

[108] Robert Heler, Poulami Samai, Joshua W. Modell, Catherine Weiner, Gregory W. Goldberg, David Bikard, and Luciano a. Marraffini, *Cas9 specifies functional viral targets during CRISPR Cas adaptation*, Nature **519** (2015), no. 7542, 199–202.

[109] Robert Heler, Addison V. Wright, Marija Vucelja, David Bikard, Jennifer A. Doudna, and Luciano A. Marraffini, *Mutations in Cas9 Enhance the Rate of Acquisition of Viral Spacer Sequences during the CRISPR-Cas Immune Response*, Molecular Cell **65** (2016), no. 1, 168–175.

[110] Gary E. Heussler, Jon L. Miller, Courtney E. Price, Alan J. Collins, and George A. O'Toole, *Requirements for Pseudomonas aeruginosa Type I-F CRISPR-Cas Adaptation Determined Using a Biofilm Enrichment Assay*, Journal of Bacteriology **198** (2016), no. 22, JB.00458–16.

[111] Gary E. Heussler and George A. O'Toole, *Friendly Fire: Biological Functions and Consequences of Chromosomal-Targeting by CRISPR-Cas Systems*, Journal of Bacteriology **198** (2016), no. 10, 1481–1486.

[112] Alison B Hickman and Fred Dyda, *The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications*, Nucleic Acids Research **43** (2015), no. 16, 1–12.

[113] Steven P T Hooton and Ian F Connerton, *Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein.*, Frontiers in microbiology **5** (2014), no. January, 744.

[114] Hannes Horn, Beate Slaby, Martin Jahn, Kristina Bayer, Lucas Moitinhosilva, Frank Förster, Usama R Abdelmohsen, and Ute Hentschel, *An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes*, Frontiers in M **7** (2016), no. 1751.

[115] Stineke Van Houte, Angus Buckling, and Edze R Westra, *Immigration of susceptible hosts triggers the evolution of alternative parasite defence strategies.*, Proceedings of the Royal Society B: Biological Sciences **283** (2016), no. 1837.

[116] Nina M Høyland-Kroghsbo, Jon Paczkowski, Sampriti Mukherjee, Jenny Broniewski, Edze Westra, Joseph Bondy-Denomy, and Bonnie L Bassler, *Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system.*, Proceedings of the National Academy of Sciences of the United States of America **114** (2016), no. 1, 131–135.

[117] Qinqin Huang, Hongping Luo, Minqiang Liu, Jie Zeng, Abualgasim Elgaili Abdalla, Xiangke Duan, Qiming Li, and Jianping Xie, *The effect of Mycobacterium tuberculosis CRISPR-associated Cas2 (Rv2816c) on stress response genes expression, morphology and macrophage survival of Mycobacterium smegmatis*, Infection, Genetics and Evolution **40** (2015), 295–301.

[118] Alexander P Hynes, ve M Rousseau, Marie-Laurence Lemay, Philippe Horvath, Dennis A Romero, Christophe Fremaux, and Sylvain Moineau, *An*

anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9, Nature Microbiology **Epub ahead** (2017).

[119] Alexander P. Hynes, Manuela Villion, and Sylvain Moineau, *Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages*, Nature Communications **5** (2014), no. 4399.

[120] Jaime Iranzo, Alexander E Lobkovsky, Yuri I Wolf, and Eugene V Koonin, *Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context.*, Journal of bacteriology **195** (2013), no. 17, 3834–44.

[121] 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 **169** (1987), no. 12, 5429–33.

[122] Ivana Ivančić-Baće, Simon D. Cass, Stephen J. Wearne, and Edward L. Bolt, *Different genome stability proteins underpin primed and naïve adaptation in E. coli CRISPR-Cas immunity.*, Nucleic Acids Research **43** (2015), no. 22, 10821–30.

[123] Ryan N. Jackson, Sarah M. Golden, Paul B. G. van Erp, Joshua Carter, Edze R. Westra, Stan J. J. Brouns, John van der Oost, Thomas C. Terwilliger, Randy J. Read, and Blake Wiedenheft, *Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli*, Science **345** (2014), no. 6203, 1479–84.

[124] Simon A Jackson, Rebecca E McKenzie, Robert D Fagerlund, Sebastian N Kieper, Peter C Fineran, and Stan J J Brouns, *CRISPR-Cas: Adapting to change*, Science **356** (2017), no. 6333, eaal5056.

[125] Rund Jansen, Jam D.A. van Embden, Wim Gaastra, and Leo M. Schouls, *Identification of a Novel Family of Sequence Repeats among Prokaryotes*, OMICS: A Journal of Integrative Biology **6** (2002), no. 1, 23–33.

[126] Ruud Jansen, Jan D A Van Embden, Wim Gaastra, and Leo M. Schouls, *Identification of genes that are associated with DNA repeats in prokaryotes*, Molecular Microbiology **43** (2002), no. 6, 1565–1575.

[127] Wenyan Jiang, David Bikard, David Cox, Feng Zhang, and Luciano a Marraffini, *RNA-guided editing of bacterial genomes using CRISPR-Cas systems.*, Nature biotechnology **31** (2013), no. 3, 233–9.

[128] Wenyan Jiang, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R Levin, and Luciano a Marraffini, *Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids.*, PLoS genetics **9** (2013), no. 9, e1003844.

[129] Wenyan Jiang, Poulami Samai, and Luciano A. Marraffini, *Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity*, Cell **164** (2016), no. 4, 710–721.

[130] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer a Doudna, and Emmanuelle Charpentier, *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.*, Science **337** (2012), no. 6096, 816–21.

[131] Martin Jinek, Fuguo Jiang, David W Taylor, Samuel H Sternberg, Emine Kaya, Enbo Ma, Carolin Anders, Michael Hauer, Kaihong Zhou, Steven Lin, Matias Kaplan, Anthony T Iavarone, Emmanuelle Charpentier, Eva Nogales, and Jennifer A Doudna, *Structures of Cas9 endonucleases reveal RNA-mediated conformational activation.*, Science **343** (2014), no. 6176, 1247997.

[132] Matthijs M Jore, Magnus Lundgren, Esther Van Duijn, Jelle B Bultema, Edze R Westra, Sakharam P Waghmare, Blake Wiedenheft, Ümit Pul, Reinhild Wurm, Rolf Wagner, Marieke R Beijer, Arjan Barendregt, Kaihong Zhou, Ambrosius P L Snijders, Mark J Dickman, Jennifer A Doudna, Egbert J Boekema, Albert J R Heck, John Van Der Oost, and Stan J J Brouns, *Structural basis for CRISPR RNA-guided DNA recognition by Cascade.*, Nature Structural & Molecular Biology **18** (2011), no. 5, 529–536.

[133] Peter Jorth and Marvin Whiteley, *An evolutionary link between natural transformation and crispr adaptive immunity*, mBio **3** (2012), no. 5, 1–7.

[134] Donghyun Ka, Hasup Lee, Yi-Deun Jung, Kyunggon Kim, Chaok Seok, Nayoung Suh, and Euiyoung Bae, *Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System*, Structure **24** (2015), no. 1, 70–79.

[135] Vladimir V. Kapitonov, Kira S. Makarova, and Eugene V. Koonin, *ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs*, Journal of Bacteriology **198** (2015), no. 5, 797–807.

[136] Migle Kazlauskiene, Georgij Kostiuk, Česlovas Venclovas, Gintautas Tamulaitis, and Virginijus Siksnys, *A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems.*, Science **357** (2017), no. 6351, 605–609.

[137] Tom Killelea and Edward L Bolt, *CRISPR-Cas Adaptive Immunity and the Three Rs*, Bioscience Reports **37** (2017), no. 4.

[138] Yoon Koo, Du Kyo Jung, and Euiyoung Bae, *Crystal structure of streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure*, PLoS ONE **7** (2012), no. 3, e33401.

[139] Eugene V Koonin and Mart Krupovic, *Evolution of adaptive immunity from transposable elements combined with innate immune systems.*, Nature Reviews Genetics **16** (2014), no. 3, 184–92.

[140] Eugene V Koonin, Kira S Makarova, and Feng Zhang, *Diversity, classification and evolution of CRISPR-Cas systems*, Current Opinion in Microbiology **37** (2017), 67–78.

[141] Mart Krupovic, Pierre Béguin, and Eugene V. Koonin, *Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery*, Current Opinion in Microbiology **38** (2017), 36–43.

[142] Mart Krupovic, Kira S Makarova, Patrick Forterre, David Prangishvili, and Eugene V Koonin, *Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity.*, BMC biology **12** (2014), no. 1, 36.

[143] Mart Krupovic, Sergey Shmakov, Kira S. Makarova, Patrick Forterre, and Eugene V. Koonin, *Recent mobility of casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity*, Genome Biology and Evolution **8** (2016), no. 2, 375–386.

[144] Tim Kü, Sebastian N Kieper, Jasper W Bannenberg, Martin Depken, Maria Suarez-Diez, Stan J J Brouns, Tim Kü Nne, Anne I M Vogel, Willem R Miellet, and Misha Klein, *Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation*, Molecular Cell **63** (2016), no. 5, 852–864.

[145] Anne Kupczok, Giddy Landan, and Tal Dagan, *The contribution of genetic recombination to CRISPR array evolution*, Genome Biology and Evolution **7** (2015), no. 7, 1925–1939.

[146] Eric S Lander, *The Heroes of CRISPR*, Cell **164** (2015), no. 1-2, 18–28.

[147] Kwang Hoon Lee, Seong Gyu Lee, Kyung Eun Lee, Hyesung Jeon, Howard Robinson, and Byung Ha Oh, *Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity*, Proteins: Structure, Function and Bioinformatics **80** (2012), no. 11, 2573–2582.

[148] Sofia Lemak, Natalia Beloglazova, Boguslaw Nocek, Tatiana Skarina, Robert Flick, Greg Brown, Ana Popovic, Andrzej Joachimiak, Alexei Savchenko, and Alexander F. Yakunin, *Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus*, Journal of the American Chemical Society **135** (2013), no. 46, 17476–17487.

[149] Asaf Levy, Moran G Goren, Ido Yosef, Oren Auster, Miriam Manor, Gil Amitai, Rotem Edgar, Udi Qimron, and Rotem Sorek, *CRISPR adaptation biases explain preference for acquisition of foreign DNA*, Nature **520** (2015), no. 7548, 505–510.

[150] Ming Li, Rui Wang, Dahe Zhao, and Hua Xiang, *Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process*, Nucleic Acids Research **42** (2014), no. 4, 2483–2492.

[151] Tzu-Lung Lin, Yi-Jiun Pan, Pei-Fang Hsieh, Chun-Ru Hsu, Meng-Chuan Wu, and Jin-Town Wang, *Imipenem represses CRISPR-Cas interference of DNA acquisition through H-NS stimulation in Klebsiella pneumoniae.*, Scientific reports **6** (2016), 31644.

[152] Tao Liu, Zhenzhen Liu, Qing Ye, Saifu Pan, Xiaodi Wang, Yingjun Li, Wenfang Peng, Yunxiang Liang, Qunxin She, and Nan Peng, *Coupling transcriptional activation of CRISPR Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus*, Nucleic Acids Research **45** (2017), no. 15, 8978–8992.

[153] R G Lloyd and G J Sharples, *Dissociation of synthetic Holliday junctions by E. coli RecG protein.*, The EMBO journal **12** (1993), no. 1, 17–22.

[154] Anna Lopatina, Sofia Medvedeva, Sergey Shmakov, Maria D. Logacheva, Vjacheslav Krylenkov, and Konstantin Severinov, *Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow*, Frontiers in Microbiology **7** (2016), 1–13.

[155] Maria José Lopez-Sanchez, Elisabeth Sauvage, Violette Da Cunha, Dominique Clermont, Elisoa Ratsima Hariniaina, Bruno Gonzalez-Zorn, Claire Poyart, Isabelle Rosinski-Chupin, and Philippe Glaser, *The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome*, Molecular Microbiology **85** (2012), no. 6, 1057–1071.

[156] R. Louwen, D. Horst-Kreft, A. G. De Boer, L. Van Der Graaf, G. De Knegt, M. Hamersma, A. P. Heikema, A. R. Timms, B. C. Jacobs, J. A. Wagenaar, H. P. Endtz, J. Van Der Oost, J. M. Wells, E. E S Nieuwenhuis, A. H M Van Vliet, P. T J Willemsen, P. Van Baarlen, and A. Van Belkum, *A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barré syndrome*, European Journal of Clinical Microbiology and Infectious Diseases **32** (2013), no. 2, 207–226.

[157] Guoqin Mai, Ruiquan Ge, Guoquan Sun, Qinghan Meng, and Fengfeng Zhou, *A Comprehensive Curation Shows the Dynamic Evolutionary Patterns of Prokaryotic CRISPRs*, BioMed Research International **Epub** (2016).

[158] Sonali Majumdar, Peng Zhao, Neil T Pfister, Mark Compton, Sara Olson, Claiborne V C Glover, Lance Wells, Brenton R Graveley, Rebecca M Terns,

and Michael P Terns, *Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus.*, RNA **21** (2015), no. 6, 1147–58.

[159] Kira S. Makarova, *A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis*, Nucleic Acids Research **30** (2002), no. 2, 482–496.

[160] Kira S Makarova, Nick V Grishin, Svetlana a Shabalina, Yuri I Wolf, and Eugene V Koonin, *A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.*, Biology direct **1** (2006).

[161] Kira S. Makarova, Yuri I. Wolf, Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen, and Eugene V. Koonin, *An updated evolutionary classification of CRISPR Cas systems*, Nature Reviews Microbiology **13** (2015), no. 11, 722–736.

[162] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin, *The basic building blocks and evolution of CRISPR-Cas systems.*, Biochemical Society transactions **41** (2013), no. 6, 1392–400.

[163] Kira S. Makarova, Feng Zhang, and Eugene V. Koonin, *SnapShot: Class 1 CRISPR-Cas Systems*, Cell **168** (2017), no. 5.

[164] ———, *SnapShot: Class 2 CRISPR-Cas Systems*, Cell **168** (2017), no. 1-2.

[165] Prashant Mali, Kevin M Esvelt, and George M Church, *Cas9 as a versatile tool for engineering biology.*, Nature methods **10** (2013), no. 10, 957–63.

[166] Pierre Mandin, Thomas Geissmann, Pascale Cossart, Francis Repoila, and Massimo Vergassola, *Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets*, Nucleic Acids Research **35** (2007), no. 3, 962–974.

[167] Tatiana C. Mangericao, Zhanhao Peng, and Xuegong Zhang, *Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls*, BMC Systems Biology **10** (2016), no. S1, 5.

[168] Luciano A. Marraffini, *CRISPR-Cas immunity in prokaryotes*, Nature **526** (2015), no. 7571, 55–61.

[169] Luciano A Marraffini and Erik J Sontheimer, *CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.*, Science **322** (2008), no. 5909, 1843–1845.

[170] Luciano a Marraffini and Erik J Sontheimer, *Self versus non-self discrimination during CRISPR RNA-directed immunity.*, Nature **463** (2010), no. 7280, 568–71.

[171] Judita Mascarenhas, Humberto Sanchez, Serkalem Tadesse, Dawit Kidane, Mahalakshmi Krisnamurthy, Juan C Alonso, and Peter L Graumann, *Bacillus subtilis SbcC protein plays an important role in DNA inter-strand cross-link repair.*, BMC molecular biology **7** (2006), 20.

[172] Andreas Mayer, Thierry Mora, Olivier Rivoire, and Aleksandra M Walczak, *Diversity of immune strategies explained by adaptation to pathogen statistics*, Proceedings of the National Academy of Sciences **113** (2015), no. 31, 8630–8635.

[173] Jon McGinn and Luciano A. Marraffini, *CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration*, Molecular Cell **64** (2016), no. 3, 616–623.

[174] L. Medina-Aparicio, J. E. Rebollar-Flores, A. L. Gallego-Hernández, A. Vázquez, L. Olvera, R. M. Gutiérrez-Ríos, E. Calva, and I. Hernández-Lucas, *The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in Salmonella enterica serovar Typhi*, Journal of Bacteriology **193** (2011), no. 10, 2396–2407.

[175] Bénédicte Michel and David Leach, *Homologous Recombinatio Enzymes and Pathways*, EcoSal Plus **5** (2012), no. 1.

[176] Eran Mick, Adi Stern, and Rotem Sorek, *Holding a grudge*, RNA Biology **10** (2013), no. 5, 900–906.

[177] Anne M. Millen, Philippe Horvath, Patrick Boyaval, and Dennis a. Romero, *Mobile CRISPR/Cas-Mediated Bacteriophage Resistance in Lactococcus lactis*, PLoS ONE **7** (2012), no. 12, e51663.

[178] Samuel Minot, Rohini Sinha, Jun Chen, Hongzhe Li, Sue a Keilbaugh, Gary D Wu, James D Lewis, and Frederic D Bushman, *The human gut virome : Inter-individual variation and dynamic response to diet*, Genome research **21** (2011), no. 10, 1616–1625.

[179] Joshua W. Modell, Wenyan Jiang, and Luciano A. Marraffini, *CRISPR Cas systems exploit viral DNA injection to establish and maintain adaptive immunity*, Nature **544** (2017), no. 7648, 101–104.

[180] Ralf Moeller, Erko Stackebrandt, Günther Reitz, Thomas Berger, Petra Rettberg, Aidan J. Doherty, Gerda Horneck, and Wayne L. Nicholson, *Role of DNA repair by nonhomologous-end joining in Bacillus subtilis spore resistance to extreme dryness, mono- and polychromatic UV, and ionizing radiation*, Journal of Bacteriology **189** (2007), no. 8, 3306–3311.

[181] P. Mohanraju, K.S. Makarova, B. Zetsche, F. Zhang, E.V. Koonin, and J Van der Oost, *Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems*, Science **353** (2016), no. 6299, aad5147.

[182] Francisco J.M. Mojica, César Díez-Villaseñor, Jesús García-Martínez, and Elena Soria, *Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements*, Journal of Molecular Evolution **60** (2005), no. 2, 174–182.

[183] Francisco J.M. Mojica, C. Ferrer, Guadalupe. Juez, and Francisco Rodríguez-Valera, *Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning*, Molecular Microbiology **17** (1995), no. 1, 85–93.

[184] Daniel Morley, Jenny M. Broniewski, Edze R. Westra, Angus Buckling, and Stineke van Houte, *Host diversity limits the evolution of parasite local adaptation*, Molecular Ecology **26** (2017), no. 7, 1756–1763.

[185] Ioannis Mougiakos, Elleke F Bosma, Willem M de Vos, Richard van Kranenburg, and John van der Oost, *Next Generation Prokaryotic Engineering: The CRISPR-Cas Toolkit*, Trends in Biotechnology **34** (2016), no. 7, 575–587.

[186] Sabin Mulepati, Annie Héroux, and Scott Bailey, *Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target.*, Science **345** (2014), no. 6203, 1479–84.

[187] Ki Hyun Nam, Charles Haitjema, Xueqi Liu, Fran Ding, Hongwei Wang, Matthew P. Delisa, and Ailong Ke, *Cas5d protein processes Pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg crispr-cas system*, Structure **20** (2012), no. 9, 1574–1584.

[188] Saskia B. Neher, Judit Villén, Elizabeth C. Oakes, Corey E. Bakalarski, Robert T. Sauer, Steven P. Gygi, and Tania A. Baker, *Proteomic Profiling of ClpXP Substrates after DNA Damage Reveals Extensive Instability within SOS Regulon*, Molecular Cell **22** (2006), no. 2, 193–204.

[189] Ole Niewoehner, Carmela Garcia-doval, Jakob T Rostøl, Christian Berk, Frank Schwede, Laurent Bigler, Jonathan Hall, Luciano A Marraffini, and Martin Jinek, *Type III CRISPR-Cas systems generate cyclic oligoadenylate*

second messengers to activate Csm6 RNases, Nature **548** (2017), no. 7669, 543–548.

[190] Ole Niewoehner, Martin Jinek, and Jennifer A. Doudna, *Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases*, Nucleic Acids Research **42** (2014), no. 2, 1341–1353.

[191] Hiroshi Nishimasu, F. Ann Ran, Patrick D. Hsu, Silvana Konermann, Soraya I. Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki, *Crystal structure of Cas9 in complex with guide RNA and target DNA*, Cell **156** (2014), no. 5, 935–949.

[192] James K. Nuñez, Lawrence Bai, Lucas B. Harrington, Tracey L. Hinder, and Jennifer A. Doudna, *CRISPR Immunological Memory Requires a Host Factor for Specificity*, Molecular Cell **62** (2016), no. 6, 824–833.

[193] James K. Nuñez, Lucas B. Harrington, Philip J. Kranzusch, Alan N. Engelman, and Jennifer A. Doudna, *Foreign DNA capture during CRISPRCas adaptive immunity*, Nature **527** (2015), no. 7579, 535–538.

[194] James K. Nuñez, Amy S. Y. Lee, Alan Engelman, and Jennifer a. Doudna, *Integrase mediated spacer acquisition during CRISPR Cas adaptive immunity*, Nature **519** (2015), no. 7542, 193–198.

[195] Jee Hwan Oh and Jan Peter Van Pijkeren, *CRISPR-Cas9-assisted recombineering in Lactobacillus reuteri*, Nucleic Acids Research **42** (2014), no. 17, 1–11.

[196] Pedro H Oliveira, Marie Touchon, and Eduardo P C Rocha, *The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts.*, Nucleic acids research **42** (2014), no. 16, 10618–10632.

[197] Pedro H Oliveira, Marie Touchon, and Eduardo P.C. Rocha, *Regulation of genetic flux between bacteria by restriction modification systems*, Proceedings of the National Academy of Sciences of the United States of America **113** (2016), no. 20, 5658–5663.

[198] David Paez-Espino, Wesley Morovic, Christine L Sun, Brian C Thomas, Kenichi Ueda, Buffy Stahl, Rodolphe Barrangou, and Jillian F Banfield, *Strong bias in the bacterial CRISPR elements that confer immunity to phage.*, Nature communications **4** (2013), 1430.

[199] David Paez-espino, Itai Sharon, Wesley Morovic, Buffy Stahl, Brian C Thomas, Rodolphe Barrangou, and F Banfield, *CRISPR Immunity Drives Rapid Phage Genome Evolution in Streptococcus thermophilus*, mBio **6** (2015), no. 2, 1–9.

[200] M. Pagel, *Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters*, Proceedings of the Royal Society B: Biological Sciences **255** (1994), no. 1342, 37–45.

[201] Mark Pagel and Andrew Meade, *Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible Jump Markov Chain Monte Carlo.*, The American Naturalist **167** (2013), no. 6, 808–825.

[202] Kelli L Palmer and Michael S. Gilmore, *Multidrug-Resistant Enterococci Lack CRISPR cas*, mBio **1** (2010), no. 4, 1–10.

[203] Ülvi Paris, Katren Mikkel, Kairi Tavita, Signe Saumaa, Riho Teras, and Maia Kivisaar, *NHEJ enzymes LigD and Ku participate in stationary-phase mutagenesis in Pseudomonas putida*, DNA Repair **31** (2015), 11–18.

[204] Adrian G. Patterson, James T. Chang, Corinda Taylor, and Peter C. Fineran, *Regulation of the type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference*, Nucleic Acids Research **43** (2015), no. 12, 6038–6048.

[205] Adrian G. Patterson, Simon A. Jackson, Corinda Taylor, Gary B. Evans, George P.C. Salmond, Rita Przybilski, Raymond H.J. Staals, and Peter C. Fineran, *Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems*, Molecular Cell **64** (2016), no. 6, 1–7.

[206] Adrian G Patterson, Mariya S Yevstigneyeva, and C Peter, *Regulation of CRISPRCas adaptive immune systems*, Current Opinion in Microbiology **37** (2017), 1–7.

[207] April Pawluk, Nadia Amrani, Yan Zhang, Bianca Garcia, Yurima Hidalgo-Reyes, Jooyoung Lee, Alireza Edraki, Megha Shah, Erik J. Sontheimer, Karen L. Maxwell, and Alan R. Davidson, *Naturally Occurring Off-Switches for CRISPR-Cas9*, Cell **167** (2016), no. 7, 1829–1838.

[208] April Pawluk, Joseph Bondy-Denomy, Vivian H W Cheung, Karen L. Maxwell, and Alan R. Davidson, *A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of pseudomonas aeruginosa*, mBio **5** (2014), no. 2, 1–7.

[209] April Pawluk, Raymond H.J. Staals, Corinda Taylor, Bridget N.J. Watson, Senjuti Saha, Peter C. Fineran, Karen L. Maxwell, and Alan R. Davidson, *Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species*, Nature Microbiology **1** (2016), no. 8, 16085.

[210] Bruce M. Pearson, Rogier Louwen, Peter van Baarlen, and Arnoud H.M. van Vliet, *Differential distribution of Type II CRISPR-Cas systems in agricultural and non-agricultural Campylobacter coli and Campylobacter jejuni*

isolates correlates with lack of shared environments, Genome Biology and Evolution **7** (2015), no. 9, 2663–2679.

[211] Ritsdeliz Perez-Rodriguez, Charles Haitjema, Qingqiu Huang, Ki Hyun Nam, Sarah Bernardis, Ailong Ke, and Matthew P. DeLisa, *Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli*, Molecular Microbiology **79** (2011), no. 3, 584–599.

[212] Joseph E. Peters, Kira S. Makarova, Sergey Shmakov, and Eugene V. Koonin, *Recruitment of CRISPR-Cas systems by Tn7-like transposons*, Proceedings of the National Academy of Sciences of the United States of America **114** (2017), no. 35, E7358–E7366.

[213] Robert S Pitcher, Nigel C Brissett, and Aidan J Doherty, *Non homologous end-joining in bacteria: a microbial perspective.*, Annual review of microbiology **61** (2007), 259–282.

[214] Robert S. Pitcher, Nigel C. Brissett, Angel J. Picher, Paula Andrade, Raquel Juarez, Darren Thompson, Gavin C. Fox, Luis Blanco, and Aidan J. Doherty, *Structure and Function of a Mycobacterial NHEJ DNA Repair Polymerase*, Journal of Molecular Biology **366** (2007), no. 2, 391–405.

[215] Robert S Pitcher, Andrew J Green, Anna Brzostek, Malgorzata Korycka-machala, Jaroslaw Dziadek, and Aidan J Doherty, *NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation*, DNA Repair **6** (2007), 1271–1276.

[216] André Plagens, Britta Tjaden, Anna Hagemann, Lennart Randau, and Reinhard Hensel, *Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax*, Journal of Bacteriology **194** (2012), no. 10, 2491–2500.

[217] Christine Pourcel, Gregory Salvignol, and Gilles Vergnaud, *CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies*, Microbiology **151** (2005), no. 3, 653–663.

[218] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin, *Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix*, Molecular Biology and Evolution **26** (2009), no. 7, 1641–1650.

[219] Tessa E F Quax, Marleen Voet, Odile Sismeiro, Marie-Agnes Dillies, Bernd Jagla, Jean-Yves Coppée, Guennadi Sezonov, Patrick Forterre, John van der Oost, Rob Lavigne, and David Prangishvili, *Massive activation of archaeal defense genes during viral infection.*, Journal of virology **87** (2013), no. 15, 8419–28.

[220] Nancy F. Ramia, Michael Spilman, Li Tang, Yaming Shao, Joshua Elmore, Caryn Hale, Alexis Cocozaki, Nilakshee Bhattacharya, Rebecca M. Terns, Michael P. Terns, Hong Li, and Scott M. Stagg, *Essential Structural and Functional Roles of the Cmr4 Subunit in RNA Cleavage by the Cmr CRISPR-Cas Complex*, Cell Reports **9** (2014), no. 5, 1610–1618.

[221] Chitong Rao, Cyril Guyard, Carmen Pelaz, Jessica Wasserscheid, Joseph Bondy-Denomy, Ken Dewar, and Alexander W. Ensminger, *Active and adaptive Legionella CRISPR-Cas reveals a recurrent challenge to the pathogen*, Cellular Microbiology **18** (2016), no. 10, 1319–1338.

[222] H K Ratner, T R Sampson, and D S Weiss, *I can see CRISPR now, even when phage are gone: a view on alternative CRISPR-Cas functions from the prokaryotic envelope*, Current Opinion in Infectious Diseases **28** (2015), no. 3, 267–274.

[223] Benjamin J Rauch, Melanie R Silvis, Judd F Hultquist, Christopher Waters, Michael J McGregor, Nevan J Krogan, and Joseph Bondy-Denomy, *Inhibition of CRISPR-Cas9 with Bacteriophage Proteins*, Cell **168** (2016), no. 1-2, 150–158.

[224] Sy Redding, Samuel H Sternberg, Blake Wiedenheft, A Jennifer, Eric C Greene, Sy Redding, Samuel H Sternberg, Myles Marshall, Bryan Gibb, Prashant Bhat, and Chantal K Guegler, *Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System Article Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System*, Cell **163** (2015), no. 4, 1–12.

[225] Mina Rho, Yu Wei Wu, Haixu Tang, Thomas G. Doak, and Yuzhen Ye, *Diverse CRISPRs evolving in human microbiomes*, PLoS Genetics **8** (2012), no. 6.

[226] Corinna Richter, Ron L. Dy, Rebecca E. McKenzie, Bridget N J Watson, Corinda Taylor, James T. Chang, Matthew B. McNeil, Raymond H J Staals, and Peter C. Fineran, *Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer*, Nucleic Acids Research **42** (2014), no. 13, 8516–8526.

[227] Eduardo P C Rocha, Emmanuel Cornet, and Bénédicte Michel, *Comparative and evolutionary analysis of the bacterial homologous recombination systems.*, PLoS genetics **1** (2005), no. 2, e15.

[228] Li Rongpeng, Fang Lizhu, Shirui Tan, Min Yu, Xuefeng Li, Sisi He, Yuquan Wei, Guoping Li, Jianxin Jiang, and Min Wu, *Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity*, Cell Research **26** (2016), no. 12, 1273–1287.

[229] Christophe Rouillon, Min Zhou, Jing Zhang, Argyris Politis, Victoria Beilsten-Edmands, Giuseppe Cannone, Shirley Graham, Carol V. Robinson, Laura Spagnolo, and Malcolm F. White, *Structure of the CRISPR interference complex CSM reveals key similarities with cascade*, Molecular Cell **52** (2013), no. 1, 124–134.

[230] Marius Rutkauskas, Tomas Sinkunas, Inga Songailiene, MariaS Tikhomirova, Virginijus Siksnys, and Ralf Seidel, *Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection*, Cell Reports **10** (2015), no. 9, 1534–1543.

[231] Akiko Sakai and Michael M. Cox, *RecFOR and RecOR as distinct RecA loading pathways*, Journal of Biological Chemistry **284** (2009), no. 5, 3264–3272.

[232] Poulami Samai, Nora Pyenson, Wenyan Jiang, Gregory W. Goldberg, Asma Hatoum-Aslan, and Luciano A. Marraffini, *Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity*, Cell **161** (2015), no. 5, 1164–1174.

[233] T. R. Sampson, B. a. Napier, M. R. Schroeder, R. Louwen, J. Zhao, C.-Y. Chin, H. K. Ratner, a. C. Llewellyn, C. L. Jones, H. Laroui, D. Merlin, P. Zhou, H. P. Endtz, and D. S. Weiss, *A CRISPR-Cas system enhances envelope integrity mediating antibiotic resistance and inflammasome evasion*, Proceedings of the National Academy of Sciences of the United States of America **111** (2014), no. 30, 11163–11168.

[234] Timothy R Sampson, Sunil D Saroj, Anna C Llewellyn, Yih-Ling Tzeng, and David S Weiss, *A CRISPR/Cas system mediates bacterial innate immune evasion and virulence.*, Nature **497** (2013), no. 7448, 254–7.

[235] Ingeborg Scholz, Sita J. Lange, Stephanie Hein, Wolfgang R. Hess, and Rolf Backofen, *CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein*, PLoS ONE **8** (2013), no. 2, e56470.

[236] Kimberley D Seed, David W Lazinski, Stephen B Calderwood, and Andrew Camilli, *A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity.*, Nature **494** (2013), no. 7438, 489–91.

[237] Kurt Selle, Todd R. Klaenhammer, and Rodolphe Barrangou, *CRISPR-based screening of genomic island excision events in bacteria*, Proceedings of the National Academy of Sciences of the United States of America **112** (2015), no. 26, 201508525.

[238] Ekaterina Semenova, Matthijs M Jore, Kirill a Datsenko, Anna Semenova, Edze R Westra, Barry Wanner, John van der Oost, Stan J J Brouns, and

Konstantin 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 **108** (2011), no. 25, 10098–103.

[239] M. A. Serbanescu, M. Cordova, K. Krastel, R. Flick, N. Beloglazova, A. Latos, A. F. Yakunin, D. B. Senadheera, and D. G. Cvitkovitch, *Role of the Streptococcus mutans CRISPR-Cas systems in immunity and cell physiology*, Journal of Bacteriology **197** (2015), no. 4, 749–761.

[240] Nikki Shariat and Edward G. Dudley, *CRISPRs: Molecular Signatures Used for Pathogen Subtyping*, Applied and Environmental Microbiology **80** (2014), no. 2, 430–439.

[241] Jiyung Shin, Fuguo Jiang, Jun-Jie Liu, Nicolas L Bray, Benjamin J Rauch, Seung Hyun Baik, Eva Nogales, Joseph Bondy-Denomy, Jacob E Corn, and Jennifer A Doudna, *Disabling Cas9 by an anti-CRISPR DNA mimic*, Science Advances **3** (2017), no. 7, e1701620.

[242] Sergey Shmakov, Omar O Abudayyeh, Kira S Makarova, Konstantin Severinov, Feng Zhang, Eugene V Koonin, Sergey Shmakov, Omar O Abudayyeh, Kira S Makarova, Yuri I Wolf, and Jonathan S Gootenberg, *Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems*, Molecular Cell **60** (2015), no. 3, 1–13.

[243] Sergey Shmakov, Ekaterina Savitskaya, Ekaterina Semenova, Maria D. Logacheva, Kirill A. Datsenko, and Konstantin Severinov, *Pervasive generation of oppositely oriented spacers during CRISPR adaptation*, Nucleic Acids Research **42** (2014), no. 9, 5907–5916.

[244] Sergey Shmakov, Aaron Smargon, David Scott, David Cox, Neena Pyzocha, Winston Yan, Omar O. Abudayyeh, Jonathan S. Gootenberg, Kira S. Makarova, Yuri I. Wolf, Konstantin Severinov, Feng Zhang, and Eugene V. Koonin, *Diversity and evolution of class 2 CRISPR Cas systems*, Nature Reviews Microbiology (2017).

[245] Sergey A Shmakov, Vassilii Sitnik, Kira S Makarova, Yuri I Wolf, Konstantin V Severinov, and Eugene V Koonin, *The CRISPR spacer space is dominated by sequences from the species-specific mobilome*, mBio **8** (2017), no. 5, 1–18.

[246] Stewart Shuman and Michael S Glickman, *Bacterial DNA repair by non-homologous end joining.*, Nature Reviews Microbiology **5** (2007), no. 11, 852–61.

[247] Sukrit Silas, Patricia Lucas-Elio, Simon A Jackson, Alejandra Aroca-Crevillén, Loren L Hansen, Peter C Fineran, Andrew Z Fire, and Antonio

Sánchez-Amat, *Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems*, eLife **6** (2017), e27601.

[248] Sukrit Silas, Kira S Makarova, Sergey Shmakov, David Páez-Espino, Georg Mohr, Yi Liu, Michelle Davison, Simon Roux, Siddharth R Krishnamurthy, Becky Xu, Hua Fu, Loren L Hansen, David Wang, Matthew B Sullivan, Andrew Millard, Martha R Clokie, Devaki Bhaya, Alan M Lambowitz, Nikos C Kyrpides, Eugene V Koonin, and Andrew Z Fire, *On the Origin of Reverse Transcriptase- Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires*, mBio (2017).

[249] Sukrit. Silas, Georg Mohr, David. J. Sidote, Laura. M. Markham, Antonio. Sanchez-Amat, Devaki Bhaya, Alan M. Lambowitz, and Andrew Z. Fire, *Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein*, Science **351** (2016), no. 6276, aad4234.

[250] Francisco J. Silva, Amparo Latorre, and Andres Moya, *Why are the genomes of endosymbiotic bacteria so stable?*, Trends in Genetics **19** (2003), no. 4, 172–176.

[251] Krishna Murari Sinha, Mihaela-carmen Unciuleac, Michael S Glickman, Hengyao Niu, Steven Raynard, Patrick Sung, and Stewart Shuman, *AdnAB : a new DSB-resecting motor nuclease from mycobacteria*, Genes & development **23** (2009), no. 12, 1423–1437.

[252] Erik J. Sontheimer and Alan R. Davidson, *Inhibition of CRISPR-Cas systems by mobile genetic elements*, Current Opinion in Microbiology **37** (2017), 120–127.

[253] Olga Soutourina, Marc Monot, Pierre Boudry, Laure Saujet, Christophe Pichon, Odile Sismeiro, Ekaterina Semenova, Konstantin Severinov, Chantal Le Bouguenec, Jean Yves Coppée, Bruno Dupuy, and Isabelle Martin-Verstraete, *Genome-Wide Identification of Regulatory RNAs in the Human Pathogen Clostridium difficile*, PLoS Genetics **9** (2013), no. 5.

[254] Maria Spies and Stephen C Kowalczykowski, *Homologous Recombination by the RecBCD and RecF Pathways .*, Homologous Recombination by the RecBCD and RecF Pathways, 2004.

[255] Michael Spilman, Alexis Cocozaki, Caryn Hale, Yaming Shao, Nancy Ramia, Rebeca Terns, Michael Terns, Hong Li, and Scott Stagg, *Structure of an RNA Silencing Complex of the CRISPR-Cas Immune System*, Molecular Cell **52** (2013), no. 1, 146–152.

[256] Raymond H J Staals, Yoshihiro Agari, Saori Maki-Yonekura, Yifan Zhu, David W. Taylor, Esther VanDuijn, Arjan Barendregt, Marnix Vlot,

Jasper J. Koehorst, Keiko Sakamoto, Akiko Masuda, Naoshi Dohmae, PeterJ Schaap, Jennifer A. Doudna, Albert J R Heck, Koji Yonekura, John Van der Oost, and Akeo Shinkai, *Structure and Activity of the RNA-Targeting Type III-B CRISPR-Cas Complex of Thermus thermophilus*, Molecular Cell **52** (2013), no. 1, 135–145.

[257] Raymond H J Staals, Yifan Zhu, David W. Taylor, Jack E. Kornfeld, Kundan Sharma, Arjan Barendregt, Jasper J. Koehorst, Marnix Vlot, Nirajan Neupane, Koen Varossieau, Keiko Sakamoto, Takehiro Suzuki, Naoshi Dohmae, Shigeyuki Yokoyama, Peter J. Schaap, Henning Urlaub, Albert J R Heck, Eva Nogales, Jennifer A. Doudna, Akeo Shinkai, and John VanderOost, *RNA Targeting by the Type III-A CRISPR-Cas Csm Complex of Thermus thermophilus*, Molecular Cell **56** (2014), no. 4, 518–530.

[258] Raymond H.J. Staals, Simon A. Jackson, Ambarish Biswas, Stan J.J. Brouns, Chris M Brown, Peter C. Fineran, Syouhei Nishihama, Kazuharu Yoshizuka, Xiaohong Li, and Tomonori Kawano, *Interference dominates and amplifies spacer acquisition in a native CRISPR-Cas system*, Nature Communications **23** (2016), 127–135.

[259] Adi Stern, *Self-Targeting by CRISPR : Gene regulation or autoimmunity?*, Trends in genetics : TIG **26** (2010), no. 8, 335–340.

[260] Adi Stern, Eran Mick, Itay Tirosh, Or Sagy, and Rotem Sorek, *CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome*, Genome Research **22** (2012), no. 10, 1985–1994.

[261] Samuel H. Sternberg, Benjamin LaFrance, Matias Kaplan, and Jennifer A. Doudna, *Conformational control of DNA target cleavage by CRISPRCas9*, Nature **527** (2015), no. 7576, 110–113.

[262] Samuel H Sternberg, Sy Redding, Martin Jinek, Eric C Greene, and Jennifer a Doudna, *DNA interrogation by the CRISPR RNA-guided endonuclease Cas9.*, Nature **507** (2014), no. 7490, 62–67.

[263] Christine L. Sun, Rodolphe Barrangou, Brian C. Thomas, Philippe Horvath, Christophe Fremaux, and Jillian F. Banfield, *Phage mutations in response to CRISPR diversification in a bacterial population*, Environmental Microbiology **15** (2013), no. 2, 463–470.

[264] Christine L Sun, Brian C Thomas, Rodolphe Barrangou, and Jillian F Banfield, *Metagenomic reconstructions of bacterial CRISPR loci constrain population histories*, The ISME Journal **10** (2015), no. 4, 1–13.

[265] Daan C Swarts, Cas Mosterd, Mark W J van Passel, and Stan J J Brouns, *CRISPR interference directs strand specific spacer acquisition.*, PloS one **7** (2012), no. 4, e35888.

[266] Nobuto Takeuchi, Yuri I Wolf, Kira S Makarova, and Eugene V Koonin, *Nature and intensity of selection pressure on CRISPR-associated genes.*, Journal of bacteriology **194** (2012), no. 5, 1216–25.

[267] David W Taylor, Yifan Zhu, Raymond H J Staals, and E Jack, *Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning*, Science **348** (2015), no. 6234, 581–585.

[268] Magaly Toro, Guojie Cao, Wenting Ju, Marc Allard, Rodolphe Barrangou, Shaohua Zhao, Eric Brown, and Jianghong Meng, *Association of clustered regularly interspaced short palindromic repeat (CRISPR) elements with specific serotypes and virulence potential of Shiga toxin-producing Escherichia coli*, Applied and Environmental Microbiology **80** (2014), no. 4, 1411–1420.

[269] Marie Touchon, Aude Bernheim, and Eduardo PC Rocha, *Genetic and life-history traits associated with the distribution of prophages in bacteria*, The ISME Journal **10** (2016), no. 11, 2744–2754.

[270] Marie Touchon, Sophie Charpentier, Olivier Clermont, Eduardo P C Rocha, Erick Denamur, and Catherine Branger, *CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection.*, Journal of bacteriology **193** (2011), no. 10, 2460–7.

[271] Marie Touchon, Sophie Charpentier, Dominique Pognard, Bertrand Picard, Guillaume Arlet, Eduardo P C Rocha, Erick Denamur, and Catherine Branger, *Antibiotic resistance plasmids spread among natural isolates of Escherichia coli in spite of CRISPR elements.*, Microbiology (Reading, England) **158** (2012), no. Pt 12, 2997–3004.

[272] Marie Touchon and Eduardo P C Rocha, *The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella.*, PloS one **5** (2010), no. 6, e11126.

[273] Gene W. Tyson and Jillian F. Banfield, *Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses*, Environmental Microbiology **10** (2008), no. 1, 200–207.

[274] Pedro F Vale, Guillaume Lafforgue, Francois Gatchitch, Rozenn Gardan, Sylvain Moineau, and Sylvain Gandon, *Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus*, Proceedings. Biological sciences **282** (2015), no. 1812.

[275] Paul B.G. Van Erp, Ryan N. Jackson, Joshua Carter, Sarah M. Golden, Scott Bailey, and Blake Wiedenheft, *Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli*, Nucleic Acids Research **43** (2015), no. 17, 8381–8391.

[276] Stineke van Houte, Angus Buckling, and Edze R. Westra, *Evolutionary Ecology of Prokaryotic Immune Mechanisms*, Microbiology and Molecular Biology Reviews **80** (2016), no. 3, 745–763.

[277] Stineke van Houte, Alice K.E. Ekroth, Jenny M. Broniewski, Hélène Chabas, Ben Ashby, Sylvain Gandon, Steve Paterson Mike Boots4, Angus J. Buckling, and Edze R. Westra, *The diversity-generating benefits of a prokaryotic adaptive immune system*, Nature **532** (2016), no. 7599, 385–388.

[278] Jeff L. Veesenmeyer, Aaron W. Andersen, Xiaojun Lu, Elizabeth A. Hussa, Kristen E. Murfin, John M. Chaston, Adler R. Dillman, Karen M. Wassarman, Paul W. Sternberg, and Heidi Goodrich-Blair, *NilD CRISPR RNA contributes to Xenorhabdus nematophila colonization of symbiotic host nematodes*, Molecular Microbiology **93** (2014), no. 5, 1026–1042.

[279] Reuben B. Vercoe, James T. Chang, Ron L. Dy, Corinda Taylor, Tamzin Gristwood, James S. Clulow, Corinna Richter, Rita Przybilski, Andrew R. Pitman, and Peter C. Fineran, *Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands*, PLoS Genetics **9** (2013), no. 4, e1003454.

[280] Poorna Viswanathan, Kimberly Murphy, Bryan Julien, Anthony G. Garza, and Lee Kroos, *Regulation of dev, an operon that includes genes essential for Myxococcus xanthus development and CRISPR-associated genes and repeats*, Journal of Bacteriology **189** (2007), no. 10, 3738–3750.

[281] Daria Vorontsova, Kirill A Datsenko, Sofia Medvedeva, Joseph Bondy-Denomy, Ekaterina E Savitskaya, Ksenia Pougach, Maria Logacheva, Blake Wiedenheft, Alan R Davidson, Konstantin Severinov, and Ekaterina Semenova, *Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery.*, Nucleic acids research **43** (2015), no. 22, 1–13.

[282] Jiuyu Wang, Jiazhi Li, Hongtu Zhao, Gang Sheng, Min Wang, Maolu Yin, and Yanli Wang, *Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems*, Cell **163** (2015), no. 4, 840–853.

[283] Jiuyu Wang, Jun Ma, Zhi Cheng, Xu Meng, Lilan You, Min Wang, Xinzheng Zhang, and Yanli Wang, *A CRISPR evolutionary arms race: structural insights into viral anti-CRISPR/Cas responses*, Cell Research **26** (2016), no. 10, 1165–1168.

[284] Rui Wang, Ming Li, Luyao Gong, Songnian Hu, and Hua Xiang, *DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica*, Nucleic Acids Research **44** (2016), no. 9, 4266–4277.

[285] Xiaofei Wang, Deqiang Yao, Jin-Gen Xu, A-Rong Li, Jianpo Xu, Panhan Fu, Yan Zhou, and Yongqun Zhu, *Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3*, Nature Structural & Molecular Biology **23** (2016), no. 9, 868–871.

[286] Richard A. J. Warren, *Modified Bases in Bacteriophage DNAs*, Annual Review of Microbiology **34** (1980), no. 1, 137–158.

[287] Yunzhou Wei, Megan T Chesne, Rebecca M Terns, and Michael P Terns, *Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus.*, Nucleic acids research **43** (2015), no. 3, 1749–58.

[288] Ariel D. Weinberger, Christine L. Sun, Mateusz M. Pluciński, Vincent J. Denef, Brian C. Thomas, Philippe Horvath, Rodolphe Barrangou, Michael S. Gilmore, Wayne M. Getz, and Jillian F. Banfield, *Persisting viral sequences shape microbial CRISPR-based immunity*, PLoS Computational Biology **8** (2012), no. 4.

[289] Ariel D. Weinberger, Yuri I. Wolf, Alexander E. Lobkovsky, Michael S. Gilmore, and Eugene V. Koonin, *Viral diversity threshold for adaptive immunity in prokaryotes.*, mBio **3** (2012), no. 6, 1–10.

[290] Jake L Weissman, William F Fagan, and Philip L F Johnson, *Is having more than one CRISPR array adaptive ?*, bioRxiv (2017).

[291] Geoffrey R. Weller and Aidan J. Doherty, *A family of DNA repair ligases in bacteria?*, FEBS Letters **505** (2001), no. 2, 340–342.

[292] Geoffrey R Weller, Boris Kysela, Rajat Roy, Louise M Tonkin, Elizabeth Scanlan, Marina Della, Susanne Krogh Devine, Jonathan P Day, Adam Wilkinson, Fabrizio d'Adda di Fagagna, Kevin M Devine, Richard P Bowater, Penny a Jeggo, Stephen P Jackson, and Aidan J Doherty, *Identification of a DNA nonhomologous end-joining complex in bacteria.*, Science **297** (2002), no. 5587, 1686–1689.

[293] Edze R Westra, Angus Buckling, and Peter C Fineran, *CRISPR-Cas systems: beyond adaptive immunity.*, Nature Reviews Microbiology **12** (2014), no. 5, 317–26.

[294] Edze R Westra, Andrea J Dowling, Jenny M Broniewski, and Stineke van Houte, *Evolution and Ecology of CRISPR*, Annual Review of Ecology, Evolution, and Systematics **47** (2016), no. 1, 307–331.

[295] Edze R. Westra, Umit Pul, Nadja Heidrich, Matthijs M. Jore, Magnus Lundgren, Thomas Stratmann, Reinhild Wurm, Amanda Raine, Melina Mescher, Luc Van Heereveld, Marieke Mastop, E. Gerhart H Wagner, Karin Schnetz,

John Van Der Oost, Rolf Wagner, and Stan 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 **77** (2010), no. 6, 1380–1393.

[296] Edze R. Westra, Stineke van Houte, Sam Oyesiku-Blakemore, Ben Makin, Jenny M. Broniewski, Alex Best, Joseph Bondy-Denomy, Alan Davidson, Mike Boots, and Angus Buckling, *Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense*, Current Biology **25** (2015), no. 8, 1043–1049.

[297] Matthew C Whitby and Robert G Lloyd, *Branch migration of three-strand recombination intermediates by RecG, a possible pathway for securing exchanges initiated by 3'-tailed duplex DNA.*, The EMBO journal **14** (1995), no. 14, 3302–3310.

[298] Blake Wiedenheft, Gabriel C. Lander, Kaihong Zhou, Matthijs M. Jore, Stan J. J. Brouns, John van der Oost, Jennifer A. Doudna, and Eva Nogales, *Structures of the RNA-guided surveillance complex from a bacterial immune system.*, Nature **477** (2011), no. 7365, 486–489.

[299] Dale B Wigley, *Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB.*, Nature Reviews Microbiology **11** (2013), no. 1, 9–13.

[300] Ernest Williams, Todd M. Lowe, Jeffrey Savas, and Jocelyne DiRuggiero, *Microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus exposed to gamma irradiation*, Extremophiles **11** (2007), no. 1, 19–29.

[301] Addison V Wright and Jennifer A Doudna, *Protecting genome integrity during CRISPR immune adaptation*, Nature Structural & Molecular Biology **23** (2016), no. 10.

[302] Yibei Xiao, Min Luo, Robert P Hayes, Jonathan Kim, Sherwin Ng, Fang Ding, Maofu Liao, and Ailong Ke, *Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR- Cas System*, Cell **170** (2017), no. 1, 48–60.

[303] Chaoyou Xue, Arun S. Seetharam, Olga Musharova, K Severinov, Stan J. J. Brouns, Andrew J. Severin, and Dipali G. Sashital, *CRISPR interference and priming varies with individual spacer sequences*, Nucleic acids research **43** (2015), no. 22, 10831–10847.

[304] Chaoyou Xue, Natalie R. Whitis, and Dipali G. Sashital, *Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity*, Molecular Cell **64** (2016), no. 4, 826–834.

[305] Takashi Yamano, Hiroshi Nishimasu, Bernd Zetsche, Hisato Hirano, Ian M. Slaymaker, Yinqing Li, Iana Fedorova, Takanori Nakane, Kira S. Makarova, Eugene V. Koonin, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki, *Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA*, Cell **165** (2016), no. 4, 949–962.

[306] Chaojie Yang, Peng Li, Hao Li, Hongbo Liu, Guang Yang, Jing Xie, Shengjie Yi, Jian Wang, Xianyan Cui, Zhihao Wu, Ligui Wang, Rongzhang Hao, Leili Jia, Shaofu Qiu, and Hongbin Song, *Polymorphism of CRISPR shows separated natural groupings of Shigella subtypes and evidence of horizontal transfer of CRISPR*, RNA Biology **12** (2015), no. 10, 1109–20.

[307] Hui Yang and Dinshaw J. Patel, *Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9*, Molecular Cell **67** (2017), no. 1, 117–127.

[308] K. N R Yoganand, R. Sivathanu, Siddharth Nimkar, and B. Anand, *Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system*, Nucleic Acids Research **45** (2017), no. 1, 367–381.

[309] Ido Yosef, Moran G. Goren, Ruth Kiro, Rotem Edgar, and Udi Qimron, *High-temperature protein G is essential for activity of the Escherichia coli clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system*, Proceedings of the National Academy of Sciences of the United States of America **108** (2011), no. 50, 20136–20141.

[310] Jacque C. Young, Brian D. Dill, Chongle Pan, Robert L. Hettich, Jillian F. Banfield, Manesh Shah, Christophe Fremaux, Philippe Horvath, Rodolphe Barrangou, and Nathan C. VerBerkmoes, *Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in streptococcus thermophilus*, PLoS ONE **7** (2012), no. 5, e38077.

[311] Michael E. Zegans, Jeffrey C. Wagner, Kyle C. Cady, Daniel M. Murphy, John H. Hammond, and George A. O'Toole, *Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa*, Journal of Bacteriology **91** (2009), no. 1, 210–219.

[312] Bernd Zetsche, Sara E Volz, and Feng Zhang, *A split-Cas9 architecture for inducible genome editing and transcription modulation*, Nature Biotechnology **33** (2015), no. 2, 139–142.

[313] Jing Zhang, Taciana Kasciukovic, and Malcolm F. White, *The CRISPR Associated Protein Cas4 Is a 5' to 3' DNA Exonuclease with an Iron-Sulfur Cluster*, PLoS ONE **7** (2012), no. 10, e47232.

[314] Jing Zhang, Christophe Rouillon, Melina Kerou, Judith Reeks, Kim Brugger, Shirley Graham, Julia Reimann, Giuseppe Cannone, Huanting Liu, Sonja Verena Albers, James H. Naismith, Laura Spagnolo, and Malcolm F. White, *Structure and Mechanism of the CMR Complex for CRISPR-Mediated Antiviral Immunity*, Molecular Cell **45** (2012), no. 3, 303–313.

[315] Quan Zhang, Thomas G. Doak, and Yuzhen Ye, *Expanding the catalog of cas genes with metagenomes*, Nucleic Acids Research **42** (2014), no. 4, 2448–2459.

[316] Yan Zhang, Nadja Heidrich, Biju Joseph Ampattu, Carl W Gunderson, H. Steven Seifert, Christoph Schoen, Jörg Vogel, and Erik J Sontheimer, *Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis*, Molecular Cell **50** (2013), no. 4, 488–503.

Annexe 1 : Casfinder models

Annexe 1 : Macsyfinder models for CRISPR-Cas systems

Type IA

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in TypeI
<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_I_II_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas5_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas7_TypeI" presence="mandatory" system_ref="CAS"/>

<gene name="cas7b_TypeIB" presence="accessory" system_ref="CAS-TypeIB"/>

#Specific to TypeIA
<gene name="cas5a_TypeIA" presence="mandatory"/>
<gene name="cas4_TypeIA" presence="mandatory"/>
<gene name="cas7_TypeIA" presence="mandatory"/>
<gene name="csa5_TypeIA" presence="mandatory"/>
<gene name="csaX_TypeIA" presence="mandatory"/>
<gene name="cas1_TypeIA" presence="mandatory"/>
<gene name="cas8a1a2_TypeIA" presence="mandatory"/>
<gene name="cas8a1a3_TypeIA" presence="mandatory"/>
<gene name="cas6_TypeIA" presence="mandatory"/>

</system>
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

################################
```

Type IB

```
#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in TypeI
<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_I_II_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas5_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas7_TypeI" presence="accessory"  system_ref="CAS"/>

#Specific to TypeIB
<gene name="cas5b_TypeIB" presence="mandatory"/>
<gene name="cas8a1b_TypeIB" presence="mandatory"/>
<gene name="cas7b_TypeIB" presence="mandatory"/>
<gene name="cas5b2_TypeIB" presence="mandatory"/>
<gene name="cas8b_TypeIB" presence="mandatory"/>
<gene name="cas7b2_TypeIB" presence="mandatory"/>
```

<gene name="cas8a1b2_TypeIB" presence="mandatory"/>
<gene name="cas8a1b3_TypeIB" presence="mandatory"/>
<gene name="cas7b3_TypeIB" presence="mandatory"/>

#Sometimes associated with TypeIB
<gene name="csm3_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csx10_TypeIIID" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="cas7c_TypeIC" presence="accessory" system_ref="CAS-TypeIC"/>

#Distinguishing TypeIB and TypeIC
<gene name="cas8c_TypeIC" presence="forbidden" system_ref="CAS-TypeIC"/>

</system>
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

################################

Type IC

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

<gene name="cas6_TypeI-III" presence="accessory" system_ref="CAS"/>

#Present in TypeI
<gene name="cas4_TypeI-II" presence="accessory" system_ref="CAS"/>
<gene name="cas5_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory" system_ref="CAS"/>

#Specific to TypeIC
<gene name="cas1_TypeIC" presence="mandatory"/>
<gene name="cas5c_TypeIC" presence="mandatory"/>
<gene name="cas8c_TypeIC" presence="mandatory"/>
<gene name="cas7c_TypeIC" presence="mandatory"/>
<gene name="cas5c2_TypeIC" presence="mandatory"/>
<gene name="cas7c2_TypeIC" presence="mandatory"/>
<gene name="cas8c2_TypeIC" presence="mandatory"/>

#Sometimes associated with TypeIC
<gene name="cas7b3_TypeIB" presence="accessory" system_ref="CAS-TypeIB"/>

</system>
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

################################

Type ID

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in TypeI
<gene name="cas6_TypeI-III" presence="accessory" system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory" system_ref="CAS"/>
<gene name="cas5_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory" system_ref="CAS"/>

<gene name="cas3a_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas7_TypeI" presence="accessory" system_ref="CAS"/>

#Specific to TypeID
<gene name="cas3_TypeID" presence="mandatory"/>
<gene name="csc1_TypeID" presence="mandatory"/>
<gene name="csc2_TypeID" presence="mandatory"/>
<gene name="cas10d_TypeID" presence="mandatory"/>

</system>

################################

Type IE

<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in TypeI
<gene name="cas5_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory" system_ref="CAS"/>

#Specific to TypeIE
<gene name="cas1_TypeIE" presence="mandatory"/>
<gene name="cas2_TypeIE" presence="mandatory"/>
<gene name="cas5_TypeIE" presence="mandatory"/>
<gene name="cse1_TypeIE" presence="mandatory"/>
<gene name="cse2_TypeIE" presence="mandatory"/>
<gene name="cas6_TypeIE" presence="mandatory"/>
<gene name="cas7_TypeIE" presence="mandatory"/>

#Distinguishing TypeIE from other TypeI
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory" system_ref="CAS"/>

</system>

################################

Type IF

<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in TypeI
<gene name="cas3_TypeI" presence="accessory" system_ref="CAS"/>

#Specific to TypeIF
<gene name="cas1_TypeIF" presence="mandatory"/>
<gene name="cas3-cas2_TypeIF" presence="mandatory"/>
<gene name="csy1_TypeIF" presence="mandatory"/>
<gene name="csy2_TypeIF" presence="mandatory"/>
<gene name="csy3_TypeIF" presence="mandatory"/>
<gene name="cas6_TypeIF" presence="mandatory"/>

#Distinguishing TypeIF from other TypeI
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas6_TypeI-III" presence="accessory" system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory" system_ref="CAS"/>
<gene name="cas5_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="forbidden" system_ref="CAS"/>

</system>

```
################################
```

Type I-U

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_I_II_V_maka" presence="accessory" system_ref="CAS"/>

#Present in TypeI
<gene name="cas5_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>

#Specific to TypeIU
<gene name="cas3_TypeIU" presence="mandatory"/>
<gene name="csb1_TypeIU" presence="mandatory"/>
<gene name="csb2_TypeIU" presence="mandatory"/>
<gene name="csb3_TypeIU" presence="mandatory"/>
<gene name="csx17_TypeIU" presence="mandatory"/>

#Distinguishing from other Types
<gene name="cas9_TypeII" presence="forbidden"  system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
<gene name="cas10_TypeIIIA" presence="forbidden"  system_ref="CAS-TypeIIIA"/>
<gene name="cas10_TypeIIIB" presence="forbidden"  system_ref="CAS-TypeIIIB"/>

</system>
################################
```

Type II-A

```
<system inter_gene_max_space="5" min_mandatory_genes_required="1" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in Type II
<gene name="cas1_TypeII" presence="accessory"   system_ref="CAS-TypeIIC"/>

#Specific to Type II
<gene name="cas9_TypeII" presence="accessory"   system_ref="CAS-TypeIIC" exchangeable="1">
        <homologs>
                <gene name="cas9_maka_4_II" presence="accessory"   system_ref="CAS-TypeIIC"/>
        </homologs>
</gene>

#Distinguishing from other Types
<gene name="cas3_TypeI" presence="forbidden"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="forbidden"  system_ref="CAS"/>

#Distinguishing from other TypeII
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
<gene name="csn2_TypeIIA" presence="mandatory"/>
```

<gene name="cas4_TypeI-II" presence="forbidden" system_ref="CAS"/>

</system>

################################

Type II-B

<system inter_gene_max_space="5" min_mandatory_genes_required="1" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in TypeII
<gene name="cas1_TypeII" presence="accessory"   system_ref="CAS-TypeIIC"/>

#Specific to TypeIIB
<gene name="cas4_TypeI-II" presence="mandatory" system_ref="CAS"/>
<gene name="cas9_TypeIIB" presence="mandatory" exchangeable="1">
	<homologs>
		<gene name="cas9_maka_3_IIB" presence="mandatory"/>
	</homologs>
</gene>

#Distinguishing from other Types
<gene name="cas3_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="forbidden" system_ref="CAS"/>

#Distinguishing from other TypeII
<gene name="csn2_TypeIIA" presence="forbidden" system_ref="CAS-TypeIIA"/>
<gene name="cas9_TypeII" presence="forbidden"   system_ref="CAS-TypeIIC"/>

</system>

################################

Type II-C

<system inter_gene_max_space="5" min_mandatory_genes_required="1" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>

#Present in TypeII
<gene name="cas1_TypeII" presence="accessory"/>

#Specific to TypeIIC
<gene name="cas9_TypeII" presence="mandatory" exchangeable="1">
	<homologs>
		<gene name="cas9_maka_4_II" presence="mandatory" />
	</homologs>
</gene>

#Sometimes associated with TypeIIC
<gene name="cas6_TypeI-III" presence="accessory" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory" system_ref="CAS"/>
<gene name="cas3_TypeI" presence="accessory" system_ref="CAS"/>

```
#Distinguishing from other TypeII
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
<gene name="csn2_TypeIIA" presence="forbidden"   system_ref="CAS-TypeIIA"/>
<gene name="cas4_TypeI-II" presence="forbidden"  system_ref="CAS"/>

#Distinguishing from other Types
<gene name="cas5_TypeI" presence="forbidden"  system_ref="CAS"/>

</system>

################################
```

Type III-A

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in other CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

<gene name="cas1_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas1_TypeIC" presence="accessory"  system_ref="CAS-TypeIC"/>
<gene name="cas1_TypeIE" presence="accessory"  system_ref="CAS-TypeIE"/>
<gene name="cas1_TypeIF" presence="accessory"  system_ref="CAS-TypeIF"/>
<gene name="cas1_TypeII" presence="accessory"  system_ref="CAS-TypeIIC"/>

<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>


#Specific to TypeIIIA
<gene name="cas10_TypeIIIA" presence="mandatory"/>


<gene name="csm2_TypeIIIA" presence="mandatory" exchangeable="1">
        <homologs>
                        <gene name="csm2_IIIA_maka_7" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csm3_TypeIIIA" presence="mandatory"/>

<gene name="csm4_TypeIIIA" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csm4_IIIA_maka_3" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csm5_TypeIIIA" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csm5_IIIA_maka_3" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csm6_TypeIIIA" presence="mandatory"/>


<gene name="csm2_IIIA_maka_7" presence="mandatory"/>


#Distinguishing from other Types
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas9_TypeII" presence="forbidden" system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
```

```
#Other subtypes III sometimes associated with TypeIIIA

<gene name="cas10_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr3_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr4_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr5_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr7_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr8_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>

<gene name="cas10_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr1_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr3_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr5_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr6_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>

<gene name="csm2_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIIAD_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIID_maka_6" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIID_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx19_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx10_TypeIIID" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="cas10_III_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>

<gene name="cas1_TypeII" presence="accessory" system_ref="CAS-TypeIIC"/>

</system>

################################
```

Type III-B

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in other CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

<gene name="cas1_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas1_TypeIC" presence="accessory"  system_ref="CAS-TypeIC"/>
<gene name="cas1_TypeIE" presence="accessory"  system_ref="CAS-TypeIE"/>
<gene name="cas1_TypeIF" presence="accessory"  system_ref="CAS-TypeIF"/>
<gene name="cas1_TypeII" presence="accessory"  system_ref="CAS-TypeIIC"/>
<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>

#Specific to TypeIIIB

<gene name="cmr1_IIIB_maka_4" presence="mandatory"/>
<gene name="cas10_TypeIIIB" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cas10_IIIB_maka" presence="mandatory"/>
        </homologs>
</gene>
<gene name="cmr3_TypeIIIB" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cmr3_IIIB_maka_2" presence="mandatory"/>
        </homologs>
</gene>
<gene name="cmr4_TypeIIIB" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cmr4_IIIB_maka" presence="mandatory"/>
```

```
        </homologs>
</gene>
<gene name="cmr5_TypeIIIB" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cmr5_IIIB_maka_1" presence="mandatory"/>
        </homologs>
</gene>
<gene name="cmr6_IIIB_maka_1" presence="mandatory" exchangeable="1">
        <homologs>
                        <gene name="cmr6_IIIB_maka_2" presence="mandatory"/>
        </homologs>
</gene>

#Often associated with TypeIIIB
<gene name="cmr7_IIIB_maka" presence="accessory"/>
<gene name="cmr8_IIIB_maka" presence="accessory"/>
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>

#Distinguishing from other Types
<gene name="cas9_TypeII" presence="forbidden" system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>

#Distinguishing from TypeIIIA
<gene name="cas10_TypeIIIA" presence="forbidden" system_ref="CAS-TypeIIIA"/>
<gene name="csm4_TypeIIIA" presence="forbidden" system_ref="CAS-TypeIIIA"/>
<gene name="csm5_TypeIIIA" presence="forbidden" system_ref="CAS-TypeIIIA"/>
<gene name="csm6_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>

#Other subtypes III sometimes associated with TypeIIIB

<gene name="cas10_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr1_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr3_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr5_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr6_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>

<gene name="csm2_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIIAD_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIID_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx19_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx10_TypeIIID" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="cas10_III_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>

</system>

################################
```

Type III-C

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in other CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

<gene name="cas1_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas1_TypeIC" presence="accessory"  system_ref="CAS-TypeIC"/>
<gene name="cas1_TypeIE" presence="accessory"  system_ref="CAS-TypeIE"/>
<gene name="cas1_TypeIF" presence="accessory"  system_ref="CAS-TypeIF"/>
<gene name="cas1_TypeII" presence="accessory"  system_ref="CAS-TypeIIC"/>
<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>
```

```
#Specific to TypeIIIC
<gene name="cas10_IIIC_maka" presence="mandatory"/>
<gene name="cmr1_IIIC_maka_1" presence="mandatory"/>

<gene name="cmr3_IIIC_maka" presence="mandatory" />
<gene name="cmr5_IIIC_maka_1" presence="mandatory" exchangeable="1">
<homologs>
<gene name="cmr5_IIIC_maka_2" presence="mandatory"/>
</homologs>
</gene>
<gene name="cmr6_IIIC_maka" presence="mandatory"/>

#Often associated with TypeIIIB
<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_TypeI-II" presence="accessory"  system_ref="CAS"/>

#Distinguishing from other Types
<gene name="cas9_TypeII" presence="forbidden" system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>

#Distinguishing from TypeIIIA-B-D
#Other subtypes III sometimes associated with TypeIIIB
<gene name="cas10_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm3_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm4_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm5_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm5_IIIA_maka_3" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm6_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>

<gene name="cas10_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr3_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr4_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr5_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr7_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr8_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>

<gene name="csm2_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIID_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csm3_IIIAD_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx19_IIID_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="csx10_TypeIIID" presence="accessory" system_ref="CAS-TypeIIID"/>
<gene name="cas10_III_maka_5" presence="accessory" system_ref="CAS-TypeIIID"/>

</system>

################################
```

Type III-D

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in other CRISPR-Cas systems
<gene name="cas1_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

<gene name="cas1_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas1_TypeIC" presence="accessory"  system_ref="CAS-TypeIC"/>
<gene name="cas1_TypeIE" presence="accessory"  system_ref="CAS-TypeIE"/>
<gene name="cas1_TypeIF" presence="accessory"  system_ref="CAS-TypeIF"/>
<gene name="cas1_TypeII" presence="accessory"  system_ref="CAS-TypeIIC"/>
```

```
<gene name="cas6_TypeI-III" presence="accessory"  system_ref="CAS"/>

#Specific to TypeIIID
<gene name="cas10_TypeIIIA" presence="mandatory" system_ref="CAS-TypeIIIA" exchangeable="1">
        <homologs>
                <gene name="cas10_III_maka_5" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csm2_IIID_maka_1" presence="mandatory"/>


<gene name="csx10_IIID_maka_4" presence="mandatory"/>
<gene name="csm3_IIID_maka_5" presence="mandatory"/>
<gene name="csm3_IIID_maka_6" presence="mandatory"/>

#Often associated with TypeIIID
<gene name="csm3_IIIAD_maka_1" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csm3_IIIAD_maka_5" presence="mandatory"/>
                <gene name="csm3_IIID_maka_6" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csx19_IIID_maka_1" presence="mandatory"/>

<gene name="csx10_TypeIIID" presence="mandatory"/>

<gene name="cas3_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="accessory"  system_ref="CAS"/>
<gene name="cas4_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas4_TypeI-II" presence="accessory"  system_ref="CAS"/>


#Distinguishing from other Types
<gene name="cas9_TypeII" presence="forbidden" system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>

#Other subtypes III sometimes associated with TypeIIIB
<gene name="cas10_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="cas10_III_maka_5" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm3_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm2_IIIA_maka_7" presence="accessory" system_ref="CAS-TypeIIIA"/>/>
<gene name="csm4_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm5_TypeIIIA" presence="accessory" system_ref="CAS-TypeIIIA"/>
<gene name="csm6_TypeIIIA"  presence="accessory" system_ref="CAS-TypeIIIA"/>


<gene name="cas10_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr3_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr4_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr5_TypeIIIB" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr7_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>
<gene name="cmr8_IIIB_maka" presence="accessory" system_ref="CAS-TypeIIIB"/>

<gene name="cas10_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr1_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr3_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr5_IIIC_maka_1" presence="accessory" system_ref="CAS-TypeIIIC"/>
<gene name="cmr6_IIIC_maka" presence="accessory" system_ref="CAS-TypeIIIC"/>

</system>

################################
```

Type IV

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

#Present in all CRISPR-Cas systems
<gene name="cas2_TypeI-II-III" presence="accessory"  system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>

#Specific to TypeIV
<gene name="csf1_TypeIV" presence="mandatory" exchangeable="1" >
        <homologs>
                <gene name="csf1_IV_maka_1" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csf2_TypeIV" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csf2_IV_maka_1" presence="mandatory"/>
        </homologs>
</gene>
<gene name="csf3_TypeIV" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csf3_IV_maka_1" presence="mandatory" />
        </homologs>
</gene>
<gene name="csf4_TypeIV" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="csf4_IV_maka_1" presence="mandatory"/>
        </homologs>
</gene>

<gene name="csf5_IV_maka" presence="mandatory"/>

#Associated with type IV
<gene name="cas6_TypeIF" presence="accessory" system_ref="CAS-TypeIF"/>
<gene name="cas6_TypeI-III" presence="accessory" system_ref="CAS"/>

#Distinguishing from other Types
<gene name="cas3_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas9_TypeII" presence="forbidden"  system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
<gene name="cas10_TypeIIIA" presence="forbidden"  system_ref="CAS-TypeIIIA"/>
<gene name="cas10_TypeIIIB" presence="forbidden"  system_ref="CAS-TypeIIIB"/>

</system>
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

################################
```

Type V

```
#Present in all CRISPR-Cas systems
<gene name="cas2_TypeI-II-III" presence="accessory" system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory" system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>
<gene name="cas1_TypeI-II-III" presence="accessory"   system_ref="CAS"/>

#Specific to TypeIV
<gene name="cas1_TypeV" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cas1_I_II_III_V_maka" presence="mandatory" system_ref="CAS"/>
        </homologs>
</gene>
<gene name="cas4_TypeV" presence="mandatory" exchangeable="1">
```

```
        <homologs>
                <gene name="cas4_I_II_V_maka" presence="mandatory" system_ref="CAS"/>
        </homologs>
</gene>
<gene name="cpf1_TypeV" presence="mandatory" exchangeable="1">
        <homologs>
                <gene name="cpf1_V_maka" presence="mandatory"/>
        </homologs>
</gene>

#Distinguishing from other Types
<gene name="cas3_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas3a_TypeI" presence="forbidden" system_ref="CAS"/>
<gene name="cas9_TypeII" presence="forbidden"  system_ref="CAS-TypeIIC"/>
<gene name="cas9_TypeIIB" presence="forbidden"  system_ref="CAS-TypeIIB"/>
<gene name="cas10_TypeIIIA" presence="forbidden"  system_ref="CAS-TypeIIIA"/>
<gene name="cas10_TypeIIIB" presence="forbidden"  system_ref="CAS-TypeIIIB"/>

</system>

################################
```

Type VI

```
<system inter_gene_max_space="5" min_mandatory_genes_required="1" min_genes_required="1">

<gene name="cas13a" presence="mandatory" loner="1"/>
<gene name="cas13b1" presence="mandatory" loner="1"/>
<gene name="cas13b2" presence="mandatory" loner="1"/>
<gene name="cas13c" presence="mandatory" loner="1"/>

<gene name="cas2_TypeI-II-III" presence="accessory"   system_ref="CAS"/>
<gene name="cas2_I_II_III_V_maka" presence="accessory"  system_ref="CAS"/>
<gene name="cas1_I_II_III_V_maka" presence="accessory"   system_ref="CAS"/>
<gene name="cas1_TypeI-II-III" presence="accessory"   system_ref="CAS"/>

</system>

################################
```

CAS general

```
<system inter_gene_max_space="5" min_mandatory_genes_required="2" min_genes_required="3">

<gene name="cas1_TypeI-II-III" presence="mandatory"/>
<gene name="cas1_I_II_III_V_maka" presence="mandatory"/>
<gene name="cas2_TypeI-II-III" presence="mandatory"/>
<gene name="cas2_I_II_III_V_maka" presence="mandatory"/>

<gene name="cas6_TypeI-III" presence="mandatory"/>
<gene name="cas4_TypeI-II" presence="mandatory"/>
<gene name="cas4_I_II_V_maka" presence="mandatory"/>

<gene name="cas5_TypeI" presence="accessory"/>
<gene name="cas3_TypeI" presence="accessory"/>
<gene name="cas3a_TypeI" presence="accessory"/>
<gene name="cas7_TypeI" presence="accessory"/>

<gene name="cas1_TypeIA" presence="accessory"  system_ref="CAS-TypeIA"/>
<gene name="cas1_TypeIC" presence="accessory"  system_ref="CAS-TypeIC"/>
<gene name="cas1_TypeIE" presence="accessory"  system_ref="CAS-TypeIE"/>
<gene name="cas1_TypeIF" presence="accessory"  system_ref="CAS-TypeIF"/>
<gene name="csm3_IIIAD_maka_1" presence="accessory" system_ref="CAS-TypeIIID"/>

</system>
```

Annexe 2 : Article 1 as contributing author

The following article tackles the question of the determinants of the distribution of prophages in bacterial genomes. My contribution was to perform the first detection and analysis. This initial work was used as a basis for the present article where the final analysis were carried out by Marie Touchon.

The ISME Journal (2016), 1–11

www.nature.com/ismej

ORIGINAL ARTICLE

Genetic and life-history traits associated with the distribution of prophages in bacteria

Marie Touchon[1,2], Aude Bernheim[1,2] and Eduardo PC Rocha[1,2]
[1]*Institut Pasteur, Microbial Evolutionary Genomics, Paris, France and* [2]*CNRS, UMR3525, Paris, France*

Nearly half of the sequenced bacteria are lysogens and many of their prophages encode adaptive traits. Yet, the variables driving prophage distribution remain undetermined. We identified 2246 prophages in complete bacterial genomes to study the genetic and life-history traits associated with lysogeny. While optimal growth temperatures and average cell volumes were not associated with lysogeny, prophages were more frequent in pathogens and in bacteria with small minimal doubling times. Their frequency also increased with genome size, but only for genomes smaller than 6 Mb. The number of spacers in CRISPR-Cas systems and the frequency of type III systems were anticorrelated with prophage frequency, but lysogens were more likely to encode type I and type II systems. The minimal doubling time was the trait most correlated with lysogeny, followed by genome size and pathogenicity. We propose that bacteria with highly variable growth rates often encounter lower opportunity costs for lysogeny relative to lysis. These results contribute to explain the paucity of temperate phages in certain bacterial clades and of bacterial lysogens in certain environments. They suggest that genetic and life-history traits affect the contributions of temperate phages to bacterial genomes.
The ISME Journal advance online publication, 25 March 2016; doi:10.1038/ismej.2016.47

Introduction

Temperate phages reproduce horizontally through a lytic cycle, like virulent phages, or vertically within a lysogenic host, as prophages (Lwoff, 1953). The lytic–lysogeny decision has presumably evolved from a trade-off between the relative opportunity costs of lysogeny (delayed lytic cycle) and lysis (low burst sizes under unfavorable conditions) (Weinbauer, 2004; Goldhill and Turner, 2014). In the lysogen, the interests of the prophages and their hosts are partly aligned because the former depend on the bacterium for replication. This may explain why some prophages protect the host from other phages, favor host growth or survival in certain environments, or encode toxins exploited for bacterial pathogenesis (McGrath *et al.*, 2002; Wagner and Waldor, 2002; Hyman and Abedon, 2010; Wang *et al.*, 2010). Temperate phages can thus shape the host evolution by affecting its population dynamics, through lysis, or by changing its gene repertoire, through lysogeny. They may also mediate horizontal gene transfer between bacteria (Jiang and Paul, 1998; Canchaya *et al.*, 2003a; Bobay *et al.*, 2013; Modi *et al.*, 2013).

The number of prophages in bacterial genomes is highly variable. Many bacteria are not lysogens, whereas some lysogens encode more than a dozen prophages (Fouts, 2006; Roux *et al.*, 2015). Genomic surveys showed that prophages are rare in small bacterial genomes (Casjens, 2003; Canchaya *et al.*, 2003b), where their frequency depends on the presence of restriction-modification systems (Oliveira *et al.*, 2014). To the best of our knowledge no other variables have been systematically associated with the distribution of prophages. The identification of such variables could provide new information on the genetic and life-history traits associated with lysogeny.

Environmental studies have shown that the frequency of lysogens varies in function of the environmental conditions. In particular, lysogens tend to be more abundant under conditions of low bacterial density, low nutrient concentration and low temperature (Cochran and Paul, 1998; Middelboe, 2000; Williamson *et al.*, 2002; McDaniel and Paul, 2005; Ghosh *et al.*, 2008; Pradeep Ram and Sime-Ngando, 2010; Shan *et al.*, 2014). Several arguments explain why these conditions favor lysogeny. First, they are associated with low concentrations of susceptible hosts, decreasing the benefits of lysis for the phage. Second, bacterial cells are smaller under poor growth conditions (Torrella and Morita, 1981; Akerlund *et al.*, 1995; Volkmer and Heinemann, 2011), providing fewer resources for the production of virions (reducing phage burst size). Third, prophage genes favoring host survival in poor growth conditions increase the fitness of lysogens over non-lysogens. These arguments suggest a tight association between bacterial growth conditions and lysogeny.

Correspondence: M Touchon, CNRS UMR3525, Institut Pasteur, 28 rue du Docteur Roux, Paris 75724, France.
E-mail: mtouchon@pasteur.fr
Received 27 September 2015; revised 17 February 2016; accepted 24 February 2016

The frequency of prophages depends on the outcome of a series of processes, among which the frequency of infection, the probability of lysogenizaton and the rate of prophage loss (by induction or inactivation/deletion). Several experimental studies produced a detailed picture of the molecular mechanisms underlying these processes, especially in the interaction between *Escherichia coli* and the phage Lambda (reviewed in Ptashne, 1992). Defense systems, such as CRISPR-Cas and restriction-modification systems, protect bacteria from phages (Labrie *et al*., 2010). The temperate phage that evades these defenses then faces the lytic–lysogeny decision. The frequency of lysogenization increases with the viral concentration inside the cell, which results from either high multiplicity of infection or small cell volume (Lieb, 1953; Kourilsky, 1973; Herskowitz and Hagen, 1980; St-Pierre and Endy, 2008). Finally, the rate of prophage loss by induction is higher in moments of decreased host viability, for example, following an SOS response (reviewed in Ptashne, 1992; Waldor and Friedman, 2005), under high temperatures (Bertani, 1954) or following loss of key bacterial regulators (Menouni *et al*., 2013). These studies suggest that lysogeny is associated with a multitude of traits.

Both environmental and experimental studies showed that lysogeny is favored in smaller cells and under slow growth. Bacteria able to attain very short minimal doubling times under optimal conditions (fast growers) are poorly fit to grow under poor environmental conditions (Koch, 2001). The sizes of their populations in fluctuating environments change rapidly as a consequence of oscillations between high growth rates and rapid population collapses. It has been suggested that lysogeny represents a strategy of slow replication when the host provides few resources for reproduction in waiting for more propitious conditions for productive lysis (Stewart and Levin, 1984; Abedon, 2008). In this case, lysogeny should be more frequent among fast growers because they provide more variable resources for the production of virions. Bacteria with stable growth rates provide less variable resources for phage reproduction, decreasing the potential gains of lysogeny.

Here, we wished to gain some general understanding on the variables associated with lysogeny. For this, we analyzed three variables previously highlighted by environmental and genomic studies: (1) host genome size, as previously suggested (Casjens, 2003); (2) host pathogenicity, given the numerous prophage-encoded virulence factors found in bacterial pathogens (Brussow *et al*., 2004; Abedon and Lejeune, 2005); (3) presence of CRISPR-Cas systems, given their role in defense against phages (Labrie *et al*., 2010). We also analyzed two variables highlighted by experimental studies on *E. coli*: (4) average host cell volume, since larger *E. coli* cells favor lysis over lysogeny (St-Pierre and Endy, 2008) and (5) optimal growth temperature (OGT), since high temperature favors lysis (Bertani, 1954). We added a sixth variable, directly inspired from the above-mentioned theoretical arguments on the evolution of lysogeny (Stewart and Levin, 1984; Abedon, 2008). (6) Minimal doubling times under optimal conditions, since temperate phages infecting fast growers in moments of poor growth can increase their future burst size by lysogenization.

Materials and methods

Data on bacteria

We retrieved all 2110 complete bacterial genomes of 1196 species available in Genbank (ftp://ftp.ncbi.nih.gov/genomes/, last accessed in November 2013). We extracted from primary literature and from Vieira-Silva and Rocha (2010) the minimal doubling times (d) under optimal growth condition for 223 species of bacteria. OGTs were retrieved for 222 species from the DSMZ database (http://www.dsmz.de/microorganisms/) and from Vieira-Silva and Rocha (2010)). Mesophiles were defined as organisms with OGT over 15 °C and under 60 °C. In a complementary analysis we predicted the minimal doubling times (d_{pred}) and the optimal growth temperatures ($\mathrm{OGT}_{\mathrm{pred}}$) from the genomic sequences of each of the 1196 species using Growthpred with default parameters (Vieira-Silva and Rocha, 2010). The information related with the pathogenicity of bacterial species was taken from the literature (especially Brenner *et al*., 2005).

Analyses of phages

We retrieved the complete genomes of 831 phages from Genbank Genomes (last accessed in November 2013). Temperate phages were identified using PHACTS (McNair *et al*., 2012). When the PHACTS probability score was not deemed confident we searched for the presence of integrases in phages using PFAM v26 (Finn *et al*., 2008). More specifically, we searched for proteins with significant hits to the protein profiles for tyrosine (PF00589) and serine (PF07508 and PF00239) recombinase, using HMMER3 with default options (Eddy, 2011). These predictions were manually curated using the literature and the PhAnToMe database (http://www.phantome.org).

Calculation of cell volume (V)

The volume of rods was determined from the average cell width (W) and length (L) using the formula for the volume of a cylinder capped by two hemispheres (Chrzanowski *et al*., 1988): $V = \pi (W/2)^2(L - W) + (4/3)\pi (W/2)^3$. The volume of cocci was approximated by a sphere: $V = (4/3)\pi (W/2)^3$. Length, width and shapes were retrieved from the literature (Brenner *et al*., 2005).

Detection of prophages
Prophages were detected in bacterial genomes using Phage Finder v4.6 (Fouts, 2006) (stringent option). We excluded all elements smaller than 18 kb, lacking matches to core phage proteins (e.g., terminase, capsid, head, tail proteins), or with more than 25% of insertion sequences. The latter were detected as in Touchon and Rocha (2007). Functionally related genes are usually grouped in one single region of the phage genome. Hence, elements containing several similar functional modules (e.g., integration, lysis, structural modules) more than 10 kb apart were considered as putative prophages coded in tandem. These few (~1%) elements were manually curated. Bacteria strains were considered as lysogenic when their genome contained at least one prophage. Bacterial species were defined as lysogenic when at least one strain was a lysogen.

Detection of CRISPR-Cas systems
Clusters of *cas* genes were identified and classified using MacSyFinder (Abby *et al.*, 2014). CRISPR arrays were identified following a previously published methodology (Touchon *et al.*, 2011).

Statistical analyses
Some of the variables used in this work were available for every strain (such as host genome size or the number of prophages), whereas others were only available for one or a few strains within a species (such as minimum doubling time). In 81% of the species only one complete genome was available. For the remaining species we either used all genomes (marked G in the figures) for comparisons between strain-specific traits or averaged strain-specific traits values across each species (marked S in the figures) for comparisons also involving species-specific traits. All major conclusions were controlled for the effect of phylogenetic dependency (see Supplementary Information and Supplementary Tables S1–S3). The data produced in this work is provided in Supplementary Dataset S1.

Associations between continuous variables were measured with the Spearman's rank correlation coefficient or (ρ) (Spearman, 1904). Associations between continuous and categorical variables were measured with the Wilcoxon rank-sum test (Wilcoxon, 1945). We analyzed the distribution of prophages with stepwise regressions. This standard statistical method consists in a stepwise integration of the different variables in the regression by decreasing order of contribution to the explanation of the variance of the data (Draper and Smith, 1998). We used the forward algorithm and the BIC criterion for model choice in the multiple stepwise regressions. The *P*-values associated with each variable were assessed using an *F*-test (Draper and Smith, 1998). We used JMP for the standard statistical analyses (Spearman, Wilcoxon and stepwise regressions) and the ape package in R for the analysis of phylogenetic dependency (Paradis *et al.*, 2004; see Supplementary Information).

Results

Identification and distribution of prophages in bacterial genomes
We searched for prophages in all available 2110 fully sequenced bacterial genomes (see Materials and methods). It was sometimes difficult to distinguish small partially degraded prophages from other mobile elements. Since the genomes of dsDNA self-transmissible temperate phages available in GenBank were all larger than 30 kb long, we restricted our search to prophages larger than 30 kb. We identified 2246 such elements. This constitutes our main data set of prophages. Most of these prophages encoded identifiable phage-specific functions such as integrases (86%), terminases (78%), tail- and baseplate-associated (79%), portal-associated (68%) and lysis-associated (66%) proteins. Hence, they are *bona fide* prophages.

We then searched for prophages between 18 and 30 kb long to assess how many prophage remnants or unknown small variants of intact prophages we have excluded. We identified 617 such elements. They encoded phage-specific functions at lower frequencies than in the main data set (resp. 51%, 38%, 62%, 25% and 34%), which might result from gene loss or errors in prophage identification. Unless explicitly stated otherwise, we present only the analyses made with the main data set of prophages, that is, the one including prophages larger than 30 kb. The results obtained in the analysis of the data set including smaller prophages (>18 kb) are qualitatively identical and can be found in Supplementary Material. To test if the prophages in the main data set were representative of the temperate phages present in GenBank we compared their sizes. The prophages were on average 48 kb long. This value was not significantly different from the average size of dsDNA temperate phages of GenBank (44.2 kb, see test statistics in Supplementary Figure S1). This suggests that our data set in unbiased in terms of prophage size.

Nearly half of the bacterial genomes contained at least one prophage (46% of lysogens; Figure 1a). While most lysogens had few prophages, some encoded up to 15 elements (Figure 1a). These and previous genomic (Casjens, 2003; Canchaya *et al.*, 2003b; Fouts, 2006; Roux *et al.*, 2015) and environmental analyses (Cochran and Paul, 1998; Ghosh *et al.*, 2008) suggest that lysogeny is very common in bacteria.

The effect of the host genetic background on the frequency of lysogens
The median genome size of lysogens (4.1 Mb) was twice that of non-lysogens (2.4 Mb) (Figure 1b). We tested if this difference could be justified by the

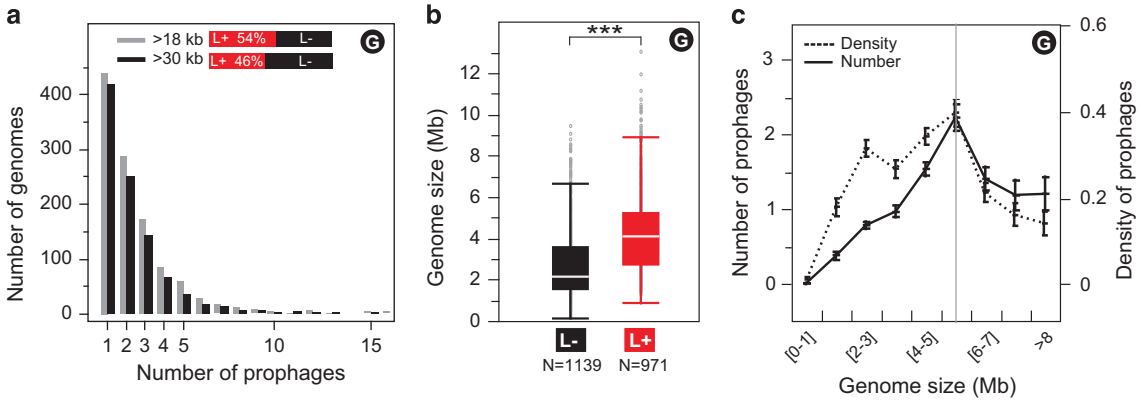


Figure 1 Distribution of prophages among all the genomes (G) used in the analysis. (**a**) Distribution of the number of prophages per genome in the two prophage data sets (>18 kb in gray, >30 kb in black). At the top: fraction of lysogens (L+) and non-lysogens (L−) in the two prophage data sets. (**b**) Box-plot of the distribution of size of the genomes (Mb) of non-lysogens (L−) and lysogens (L+) (***significant difference: $P<10^{-4}$, Wilcoxon test). The horizontal white line at the center of the box plot represents the median. The bottom and top of the box represent the first and third quartiles. The external edges of the whiskers represent the inner 10th and 90th percentiles. (**c**) Distribution of the average number (full line) and density (dash line) of prophages per host genome in function of the size of the bacterial genome (Mb) (G). The vertical gray line separates small and average from the largest bacterial genomes. There is a significant positive association between the host genome size and the number of prophages in the former (Spearman's $\rho=0.41$, $P<10^{-4}$) but not the latter (Spearman's $\rho=-0.12$, $P>0.1$). The association between the density of prophages and the host genome size is positive for the former (Spearman's $\rho=0.35$, $P<10^{-4}$) and negative for the latter (Spearman's $\rho=-0.21$, $P<10^{-4}$). Similar qualitative results were obtained in the analysis using the complementary data sets including smaller prophages and data averaged across species (Supplementary Figure S2).

increase in bacterial genome size due to prophages. Prophages accounted for an average of 3.1% of the genomes of lysogens, with a maximum of 18% in *Bartonella tribocorum* CIP 105476. These values cannot justify the median genome size difference between lysogens and non-lysogens (1.7 Mb).

The observed association between bacterial genome size and lysogeny was non-monotonic. Firstly, we found a strong positive correlation between host genome size and the number and the density of prophages in genomes up to 6 Mb (Figure 1c). This association was not exclusively caused by the absence of prophages in the small genomes of obligatory endomutualists, since it remained valid in the range 3–6 Mb (lacking obligatory endomutualists). Secondly, bacteria with genomes larger than 6 Mb, which accounted for 12% of the species in our data set, showed no significant correlation between host genome size and the number of prophages. Instead, they showed a negative correlation between host genome size and prophage density (Figure 1c). It must be noted that most of these bacteria are lysogens (77%). Overall, these results show a strong positive association between bacterial genome size and the frequency of prophages in genomes smaller than 6 Mb and no association in the largest genomes.

Smaller bacterial genomes are more compact and have fewer accessory genes. This might lead to the selection of temperate phages with smaller genomes in these hosts. This does not seem to be the case, since we found no correlation between the average size of prophages and the host genome size (Spearman's $\rho=0.01$, $P>0.8$).

We analyzed the association between CRISPR-Cas systems and lysogeny (see Materials and methods). These systems were present in 47% of the genomes, which is consistent with previous estimates (Grissa *et al.*, 2007). Intriguingly, lysogens were more likely to encode CRISPR-Cas systems (Figure 2a). Among lysogens, the number of prophages was not correlated with the presence of these systems ($P>0.6$, Wilcoxon test). Type III CRISPR-Cas systems were relatively rare in the data set (8% of all the genomes). Contrary to the general trend, bacteria encoding these specific systems carried fewer prophages and were less likely to be lysogens than the others (Figure 2b and Supplementary Figure S3).

The number of spacers in CRISPR arrays is a measure of the number of sequences targeted by the system, and presumably of its capacity to provide protection against phages. Within genomes encoding CRISPR-Cas systems, lysogens had 30% fewer CRISPR spacers than non-lysogens ($P<10^{-4}$, χ^2 test). Furthermore, we found a negative association between the number of spacers in CRISPR arrays and the number of prophages in lysogens (Figure 2c). These results show a complex association between CRISPR-Cas systems and lysogeny: lysogens tend to encode CRISPR-Cas systems with small arrays of spacers, whereas non-lysogens are more likely to either lack these systems or encode long arrays of spacers. When all the genomes were put together, there was no association between the number of CRISPR-Cas spacers and the number of prophages (Spearman's $\rho=0.04$, $P>0.1$). As a result, this variable was not used in the multivariate analyses below.

The effect of bacterial life-history traits on the frequency of lysogens

We tested the effect of bacterial life-history traits on the distribution of prophages. Most of these variables

were only available at the species level, but 19% of the species in our data set were represented by more than one genome. We averaged the strain-specific data, such as genome size and number of prophages, across species (marked S in the figures). Initially, we restricted the analysis to species with published data on bacterial cell volume (139 species), pathogenicity (668 species), OGT (222) and minimal doubling time under optimal growth conditions (223). We could complement some of these analyses with computational predictions of the traits for the remaining species (see Materials and methods).

Lysogens and non-lysogens showed no significant differences in the average cell volume (Figure 3a; see also Materials and methods). Among lysogens, we found no significant correlation between the average number of prophages carried by the genomes of a given species and the average volume of the corresponding cells after controlling for the host genome size (Supplementary Figure S4). These results show no evidence for an association between the average cell volume and lysogeny.

The OGT was not associated with lysogeny (Figure 3b, see Materials and methods). There was also no association between the average number of prophages and OGT among lysogens (Spearman's $\rho = -0.06$, $P > 0.5$). The statistical power of this analysis is weak because 202 of the 222 species with known OGT were mesophiles. We increased the size of the data set by a factor of five by predicting OGT (OGT_{pred}) for all the species. OGT can be predicted with high accuracy using protein sequences (Zeldovich *et al.*, 2007) (see Materials and methods). In this larger data set, the difference in OGT_{pred} between lysogens and non-lysogens remained non-significant when controlling for bacterial genome size (Supplementary Figure S5). Accordingly, the abundance of prophages was independent of OGT_{pred} among lysogens (Spearman's $\rho = -0.007$, $P > 0.8$).

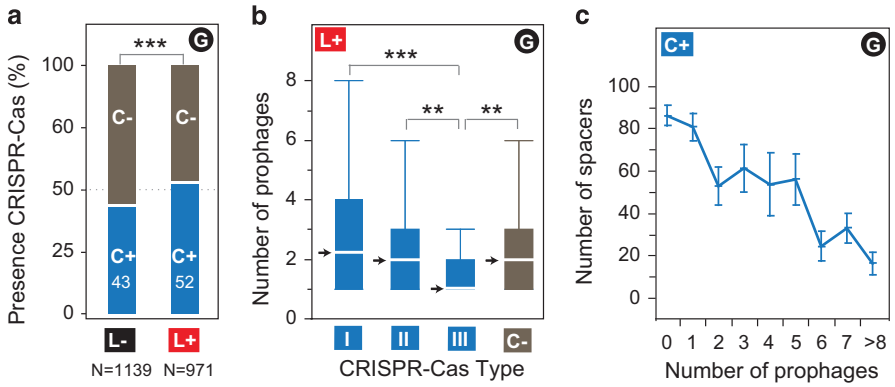


Figure 2 Analysis of the association between CRISPR-Cas systems and lysogeny among all the bacterial genomes (G). (**a**) Presence of CRISPR-Cas systems among lysogens (52%, L+) and non-lysogens (43%, L−) (***significant difference: $P < 10^{-4}$, $\chi 2$ test). (**b**) Distribution of the number of prophages per bacterial genome in lysogens (L+) in function of the presence of the different CRISPR-Cas systems (I, II, III) or when they are all absent (C−). Bacterial genomes encoding type III systems have fewer prophages than the others (***$P < 10^{-4}$ and **$P < 10^{-3}$, Wilcoxon test). Arrows indicate medians. (**c**) Distribution of the number of spacers in CRISPR arrays of bacterial genomes encoding CRISPR-Cas systems (C+) in function of the number of prophages per bacterial genome (Spearman's $\rho = -0.21$, $P < 10^{-4}$).

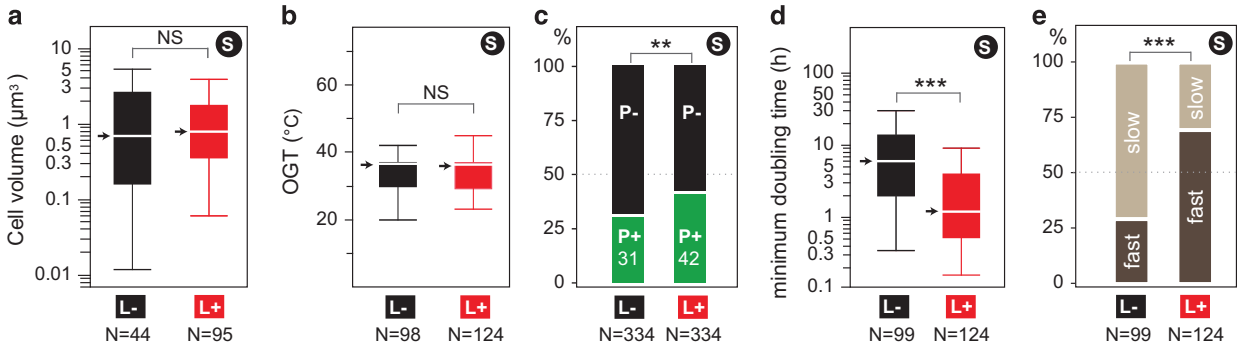


Figure 3 Analysis of the effect of species' (S) life-history traits on the distribution of lysogens. Box-plots of the distribution of the average cell volume (**a**) and optimal growth temperature (OGT, **b**) among the species with lysogens (red, L+) or lacking them (black, L−) (NS − nonsignificant differences: $P > 0.1$, Wilcoxon test). (**c**) Proportion of species including bacterial pathogens (green, P+) or lacking them (black, P−) among species with lysogens (L+) or lacking them (L−) (**significant difference: $P < 10^{-3}$; χ^2 test). Differences remained significant when controlling for genome size ($P < 10^{-4}$, stepwise regression) and phylogeny ($P < 10^{-4}$, generalized estimation equations analysis). (**d**) Box-plot of the distribution of the minimal doubling time under optimal conditions (*d*) among species with lysogens (L+) or lacking them (L−) (***significant difference: $P < 10^{-4}$, Wilcoxon test). Differences remained significant when controlling for bacterial genome size and phylogeny ($P < 10^{-4}$, generalized estimation equations analysis). (**e**) Proportion of fast (dark brown) and slow growers (light brown) among non-lysogens (L−) and lysogens (L+) (***significant difference: $P < 10^{-4}$, χ^2 test). Arrows indicate medians.

To test the association between virulence and the frequency of lysogens, we classed bacterial species into pathogens and non-pathogens (see Materials and methods). Such classifications are always coarse-grained descriptions of reality, since pathogenicity varies between strains, and depends on the eukaryotic host genetic background and physiological state. It is also difficult to class unambiguously some opportunistic bacteria (Pirofski and Casadevall, 2012). Nevertheless, species including pathogens were slightly more likely to contain prophages (see statistics in Figure 3c and Supplementary Figure S6). The observed difference might seem small, but pathogens in our data set have smaller genomes than the non-pathogens ($P<0.03$, median test). Accordingly, the frequency of prophages was higher in pathogens than in non-pathogens in all bins of genome size (see statistics in Supplementary Figure S6).

Finally, we tested the hypothesis that growth-related life-history traits affect the distribution of lysogens. We used the information on minimal doubling time under optimal conditions (d) to class bacterial species into fast growers ($d<2.5$ h) or slow growers ($d\geqslant 2.5$ h), as previously suggested (Vieira-Silva and Rocha, 2010). Strikingly, we found that the minimal doubling time of lysogens was on average five times shorter than that of non-lysogens (Figure 3d). In fact, most bacterial species with lysogens were fast growers while most others were slow growers (Figure 3e). We found a weak and nonsignificant negative correlation between the average number of prophages in lysogens and their minimal doubling time (Spearman's $\rho=-0.1$, $P>0.1$). To test these conclusions in a larger data set, we predicted the minimal doubling time of the 1196 bacterial species used in this study with Growthpred (see Materials and methods). The negative association between the minimal doubling time and the average number of prophages per host genome was highly significant in this much larger data set (Spearman's $\rho=-0.36$, $P<10^{-4}$), independently of host genome size (Supplementary Figure S7 and Supplementary Table S2).

Multivariate analysis of the variables associated with lysogeny

We found significant associations between the frequency of lysogens and host genome size, pathogenicity, and minimal doubling time. These associations were partly independent. The significant association between minimal doubling time and the average number of prophages is observed among bacterial pathogens (Spearman's $\rho=-0.48$, $P<10^{-4}$) and non-pathogens (Spearman's $\rho=-0.22$, $P<10^{-4}$; Figure 4). The associations between the frequency of lysogens and both minimal doubling time and host genome size were strictly independent. We had previously shown that minimal doubling time and genome size do not correlate (Vieira-Silva and Rocha, 2010). In the present data set slow and fast growers had similar median genome sizes (Supplementary Figure S7, both ~3.3 Mb, $P>0.8$, median test). The analysis restricted to fast growers showed that pathogenic bacteria had more prophages than the others ($P<10^{-4}$, Wilcoxon test), even if their genomes were of similar median size ($P>0.5$, median test).

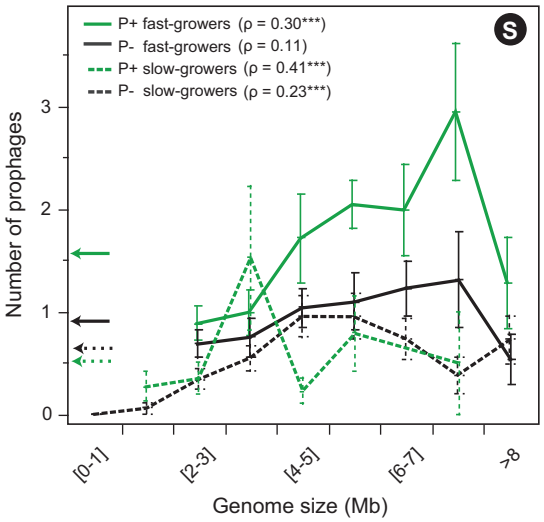


Figure 4 Distribution of the average number of prophages per bacterial genome in function of bacterial traits. The arrows on the left of the graph indicate the average number of prophages per genome (averaged across species) in each subset. The number of prophages per bacterial genome increases significantly with the host genome size in all cases (***$P<10^{-4}$, the values of Spearman's ρ are reported for each analysis), except among non-pathogenic (P−) fast growers (Spearman's $\rho=0.11$, $P>0.1$).

We used stepwise multiple regressions to test the joint effects of the three variables and to identify which variables explained more of the variance in the distribution of prophages (see Materials and methods). All three variables contributed significantly for the statistical model (BIC criterion, Supplementary Table S4). The minimal doubling time accounted for most (66%) of the explained variance, followed by host genome size (23%) and pathogenicity (11%). We extended the stepwise regression analysis to measure the interaction terms between variables, but none passed the BIC criterion.

We showed above that bacterial genome size and the frequency of lysogens were correlated only for bacterial genomes smaller than 6 Mb (Figure 1). When we restricted our regression analysis to the bacterial genomes in this range of genome size, we obtained similar results (Supplementary Table S4). In this case, the minimal doubling time accounted for 63% of the explained variance.

The stepwise regression using all the data explained a small fraction of the variance ($R^2=0.14$, $P<10^{-4}$; Supplementary Table S4). This might be due to inaccuracies in the life-history traits data to the small number of prophages per genome (that affect the statistical power of linear models), and especially to epidemiological factors increasing intra-species variance. The life-history traits (for which

phylogenetic studies are available) vary significantly only at large evolutionary scales (Galtier *et al*., 1999; Vieira-Silva *et al*., 2011). As a consequence, they might be more relevant to explain inter-species than intra-species variations in lysogeny. We tested if the inter-species variation was significant when accounting for intra-species variation, as suggested in Stearns (1977). To analyze the differences between species while reducing the effect of intra-species variation, we averaged the number of prophages per species in the set of 60 species for which there were at least five complete genomes. These species were represented by 718 genomes (34% of the data set). The stepwise regression using the 60 species showed an R^2 of 0.41 ($P<10^{-4}$; Supplementary Table S4), of which 78% was associated with the minimal doubling time. We varied the minimal number of genomes per species required to include a species in the analysis from 4 to 10 to test if this affected out conclusions. Our results show that this had little effect in the quality of the stepwise regression (Supplementary Figure S8).

The temperate phages of some bacterial phyla are poorly characterized. To test if this affected our study, we used stepwise regressions to analyze the data from Proteobacteria (which are 51% of all the bacterial genomes <6 Mb). This analysis also placed minimal doubling time as the most important explanatory variable, showing a switch in the relative order of the variables related with bacterial pathogenicity and genome size (Supplementary Table S4). Finally, we conducted the complementary analysis and removed Proteobacteria from the analysis. In this case the effects of minimal doubling time and the host genome size on the frequency of prophages remained significant (Supplementary Table S4), but the contribution of the pathogenicity was not significant. However, most large clades outside Proteobacteria had small genomes, fewer prophages and most species were non-pathogenic (Supplementary Table S5). This decreased the statistical power of the analysis.

Discussion

The traits analyzed in this work explained over 40% of the variance between species when multiple genomes were available, but seemed to explain much less of the intra-species variation. Epidemiological variables, such as the environment where the strain was isolated, might be more appropriate to model the variation of the number of prophages within species. Several other factors may have affected our results, including the accuracy of prophage detection, the biased taxonomic characteristics of the genome reference data set and the quality of the data characterizing species' traits. These problems grow in importance when species are distant from well-studied model systems. For example, one of the three variables of the stepwise regression was no longer significant when we excluded the genomes from Proteobacteria from the analysis. Nevertheless, we found qualitatively similar trends, even if quantitatively different results, in our numerous controls, which included minimal size threshold for prophages, data acquisition (literature and computed data), phylogenetic dependency and restricted range of host genome size.

We found no significant association between the frequency of lysogens and the OGT or the average cell volume. Most phages infect a relatively narrow range of hosts that have similar traits in terms of OGT and average cell volume. The lytic–lysogeny decision evolves in response to the outcomes of previous host–phage infections in this range of hosts (Hyman and Abedon, 2010). It will evolve in function of temperature and cell size deviations relative to these absolute values, not the absolute values themselves, because these deviations provide information on the relative opportunity costs of lysogeny and lysis. Previous experimental works showed that lysogeny is shaped by the variability of prokaryotic physiology (Maurice *et al*., 2013), and specifically that lysogeny is favored under suboptimal temperatures and in cells smaller than the species' average (Bertani, 1954; St-Pierre and Endy, 2008; Shan *et al*., 2014). These deviations might drive some of observed intra-species variations in lysogeny.

CRISPR-Cas systems can prevent infections by phages when they carry spacers matching their sequences. This explains why genomes encoding systems with many spacers have fewer prophages, but not why bacteria with type I and type II systems are more likely to be lysogens. Recent studies have shown a poor correlation between the presence of these CRISPR-Cas systems and the rate of horizontal gene transfer (Touchon *et al*., 2011; Gophna *et al*., 2015). If CRISPR-Cas systems with few spacers are not actively involved in immune defense against phages, as previously proposed (Touchon and Rocha, 2010; Westra *et al*., 2014), and if systems with many spacers actively protect bacteria from these elements, then our results can be reconciled with the previous experimental works: systems with long arrays prevent phage infection, resulting in few prophages in genomes, whereas the others have little impact on lysogeny.

While many lysogens encoded type I and type II CRISPR-Cas systems, very few encoded type III systems. Recent works suggested that type III-A CRISPR-Cas systems allow hosts to control their prophages (Goldberg *et al*., 2014). Phages infecting bacteria carrying these systems might have evolved to avoid lysogeny, leading to the observed negative association between lysogeny and the presence of type III systems.

We confirmed that few small bacterial genomes are lysogens. We also observed that lysogens had much larger genome sizes than would be expected given the cumulated length of the prophages they contain. Why would larger genomes have more prophages? Larger genomes are expected to have more neutral

targets for phage integration, facilitating the accumulation of these elements (Figure 5) (Bobay *et al.*, 2013). Larger genomes might directly result from the long-term accumulation of genes transferred by phages, for example, in lineages enduring frequent infections by phages. Yet, none of these hypotheses explains why this trend did not affect genomes larger than 6 Mb. If larger genomes resulted from intense selection for functional diversification by horizontal transfer, then selection for transfer might itself lead to mechanisms facilitating prophage acquisition (Cordero and Hogeweg, 2009; Smillie *et al.*, 2010). Selection for phage-related genes might saturate in the largest genomes because they contain many prophages. Alternatively, bacteria with many prophages might be very effective in preventing further phage infection (because prophages prevent infection by other phages), leading to the saturation of the number of prophages in larger genomes. Future work will be needed to quantify and disentangle the effects of host genome size on lysogeny and of lysogeny on host genome size.

We uncovered a strong negative association between minimal doubling times under optimal growth conditions and the frequency of lysogens. Minimal doubling times under optimal growth conditions and average doubling times across the diversity of conditions encountered by bacteria are not necessarily correlated (Boyce, 1984). Actually, the bacteria with the largest estimated effective population sizes are slow growers (Vieira-Silva *et al.*, 2011). The minimal doubling time is best interpreted as a key life-history trait associated with the r/K selection theory (Boyce, 1984) or with the choice between oligotrophic and copiotrophic lifestyles (Koch, 2001). Fast growers have population

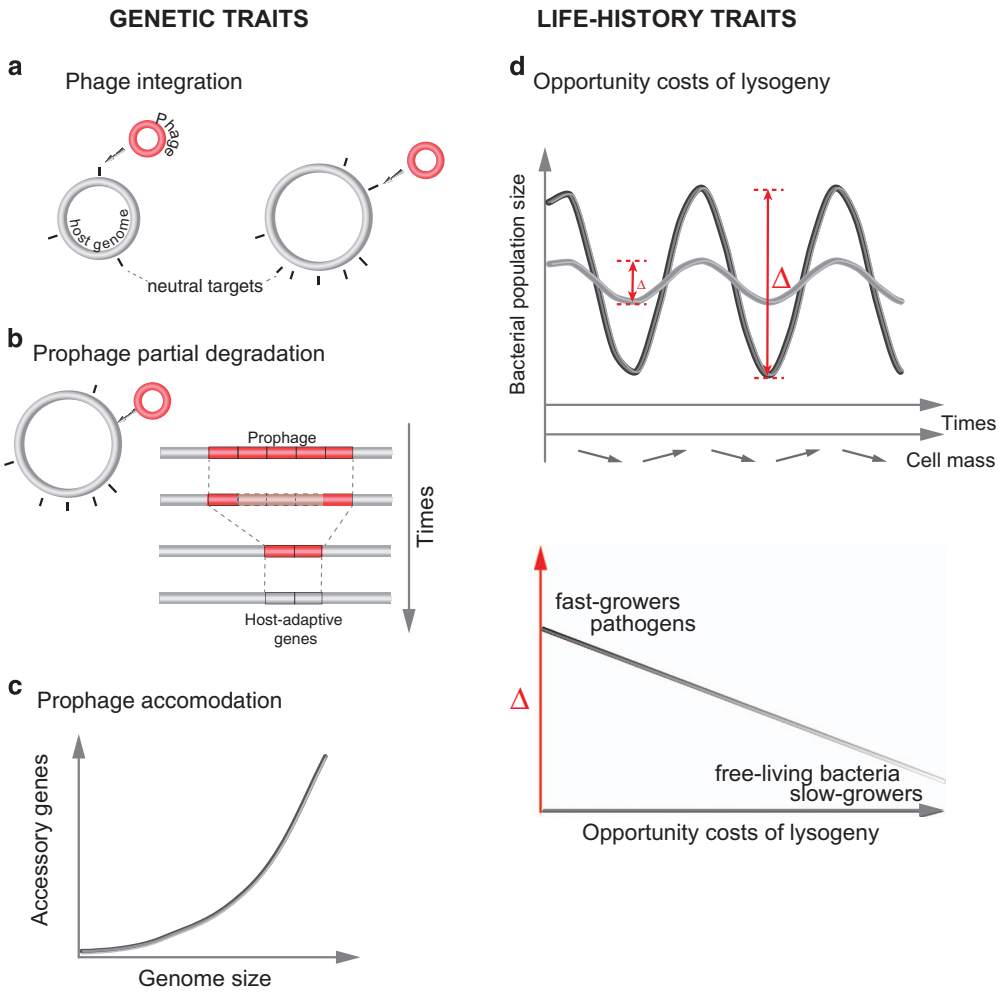


Figure 5 Genetic and life-history traits affecting the distribution of lysogens. (**a**) The number of neutral targets increases with the host genome size favoring phage integration. (**b**) Co-option of phage-related functions in degraded genetic elements increases with the number of prophages, and thus with the host genome size. After a certain time the few genes remaining in the bacterial genome may be too few or uncharacteristic to be detected as prophages. (**c**) Larger genomes have more accessory traits. (**d**) Fluctuating environmental conditions drive rapid expansion and contraction of bacterial populations (Δ), which are more important for fast growers and pathogenic bacteria than for slow growers and free-living bacteria (relative to pathogens with similar minimal doubling times). These fluctuations are associated with variations in cell mass and thus with burst size. They may also be associated with ecological conditions that constrain the lytic–lysogeny decision (such as the availability of susceptible hosts).

dynamics of alternating periods of feast and famine that are associated with large variations in growth rates and cell mass (Bremer and Dennis, 1996; Koch, 2001). The opportunity costs of lysogeny in these bacteria are very dependent on the host growth conditions at the time of infection (Figure 5). When environments are suitable, bacteria grow fast, the cell mass increases and populations are dense. This favors lytic over lysogenic cycles. Under conditions of slow bacterial growth, these populations remain at low densities and provide few resources for the production of virions; this favors lysogeny in waiting for more propitious conditions for the lytic cycle. The opportunity costs of lysogeny are generally less rewarding when phages infect slow growers because the host provides less variable resources for the production of virions. The ability to grow very fast under optimal conditions affects population dynamics (Koch, 2001), genome organization (Vieira-Silva and Rocha, 2010) and protein evolution (Vieira-Silva *et al.*, 2011). Our results suggest it also shapes the outcome of the interactions between bacteria and phages.

One could speculate that the low frequency of lysogens among slow growers could be caused by lower numbers of phages infecting these bacteria. In this case, virulent phages infecting slow-growing bacteria might also be rare. The little evidence available argues against this speculation, since many virulent phages of slow growers have been described in clades that lack lysogens in our analyses. For example, the population dynamics of cyanobacteria (slow growers and rarely lysogens) and other slow-growing marine heterotrophs are strongly affected by the numerous virulent viruses that infect them (Fuhrman, 1999; Wilhelm and Suttle, 1999; Winter *et al.*, 2010). There are also many virulent phages infecting clinical and environmental mycobacteria (Hatfull, 2010), all of which are slow growing according to our classification, but we identified few lysogens among them.

Our analyses suggest that lysogeny could be favored in bacterial pathogens. This could be explained by the virulence factors encoded by prophages (Wagner and Waldor, 2002; Brussow *et al.*, 2004), by the pathogens' peculiar cycles of population expansion and contraction (resembling those of fast growers, see above) and by the use of prophage induction as a biological weapon during colonization of a new niche (Bossi *et al.*, 2003; Gama *et al.*, 2013). The relative importance of these factors, if any, is not known.

Our work has shown associations between lysogeny and host genetic and life-history traits. These associations contribute to explain the rarity of prophages in certain clades, for example, those associated with small genomes or slow growth. Since prophages are one of the major sources of diversification of bacterial genomes, these traits may indirectly affect the evolvability of bacteria.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Louis-Marie Bobay and Mireille Ansaldi for helpful comments on earlier versions of this manuscript. This work was supported by an European Research Council starting grant (EVOMOBILOME no. 281605) to EPCR.

References

Abby SS, Neron B, Menager H, Touchon M, Rocha EP. (2014). MacSyFinder: A program to Mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One* **9**: e110726.

Abedon ST. (2008). *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Cambridge University Press: Cambridge, NY, USA.

Abedon ST, Lejeune JT. (2005). Why bacteriophage encode exotoxins and other virulence factors. *Evol Bioinform Online* **1**: 97–110.

Akerlund T, Nordstrom K, Bernander R. (1995). Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of Escherichia coli. *J Bacteriol* **177**: 6791–6797.

Bertani G. (1954). Studies on lysogenesis. III. Superinfection of lysogenic Shigella dysenteriae with temperate mutants of the carried phage. *J Bacteriol* **67**: 696–707.

Bobay L-M, Rocha EPC, Touchon M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* **30**: 737–751.

Bossi L, Fuentes JA, Mora G, Figueroa-Bossi N. (2003). Prophage contribution to bacterial population dynamics. *J Bacteriol* **185**: 6467–6471.

Boyce MS. (1984). Restitution of r-and K-selection as a model of density-dependent natural selection. *Annu Rev Ecol Syst* **15**: 427–447.

Bremer H, Dennis PP. (1996). Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt FC (ed), *Escherichia Coli and Salmonella: Cellular and Molecular Biology*. ASM Press: Washington, DC, pp 1553–1569.

Brenner DJ, Krieg NR, Staley JT. (2005). In: George M Garrity (ed), *The Proteobacteria, bergey's manual of systematic bacteriology*, 2nd edn, vol. XXVI, Springer: New York, NY, USA, 304pp. 77 illus.

Brussow H, Canchaya C, Hardt WD. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**: 560–602.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H. (2003a). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**: 417–424.

Canchaya C, Proux C, Fournoux G, Bruttin A, Brussow H. (2003b). Prophage genomics. *Microbiol Mol Biol Rev* **67**: 238–276.

Casjens S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* **49**: 277–300.

Chrzanowski TH, Crotty RD, Hubbard GJ. (1988). Seasonal variation in cell volume of epilimnetic bacteria. *Microb Ecol* **16**: 155–163.
Cochran PK, Paul JH. (1998). Seasonal abundance of lysogenic bacteria in a subtropical estuary. *Appl Environ Microbiol* **64**: 2308–2312.
Cordero OX, Hogeweg P. (2009). The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci USA* **106**: 21748–21753.
Draper NR, Smith H. (1998). *Applied Regression Analysis*. John Wiley & Sons: New York.
Eddy SR. (2011). Accelerated profile HMM searches. *PLoS Comput Biol* **7**: e1002195.
Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res* **36**: D281–D288.
Fouts DE. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**: 5839–5851.
Fuhrman JA. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.
Galtier N, Tourasse N, Gouy M. (1999). A non-hyperthermophylic common ancestor to extant life forms. *Science* **283**: 220–221.
Gama JA, Reis AM, Domingues I, Mendes-Soares H, Matos AM, Dionisio F. (2013). Temperate bacterial viruses as double-edged swords in bacterial warfare. *PLoS One* **8**: e59043.
Ghosh D, Roy K, Williamson KE, White DC, Wommack KE, Sublette KL *et al.* (2008). Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzN genes in viral-community DNA. *Appl Environ Microbiol* **74**: 495–502.
Goldberg GW, Jiang W, Bikard D, Marraffini LA. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514**: 633–637.
Goldhill DH, Turner PE. (2014). The evolution of life history trade-offs in viruses. *Curr Opin Virol* **8**: 79–84.
Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J* **9**: 2021–2027.
Grissa I, Vergnaud G, Pourcel C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.
Hatfull GF. (2010). Mycobacteriophages: genes and genomes. *Annu Rev Microbiol* **64**: 331–356.
Herskowitz I, Hagen D. (1980). The lysis-lysogeny decision of phage lambda: explicit programming and responsiveness. *Annu Rev Genet* **14**: 399–445.
Hyman P, Abedon ST. (2010). Bacteriophage host range and bacterial resistance. *Adv Appl Microbiol* **70**: 217–248.
Jiang SC, Paul JH. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol* **64**: 2780–2787.
Koch AL. (2001). Oligotrophs versus copiotrophs. *Bioessays* **23**: 657–661.
Kourilsky P. (1973). Lysogenization by bacteriophage lambda. I. Multiple infection and the lysogenic response. *Mol Gen Genet* **122**: 183–195.
Labrie SJ, Samson JE, Moineau S. (2010). Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**: 317–327.
Lieb M. (1953). Studies on lysogenization in Escherichia coli. *Cold Spring Harb Symp Quant Biol* **18**: 71–73.
Lwoff A. (1953). Lysogeny. *Bacteriol Rev* **17**: 269–337.
Maurice CF, Bouvier C, de Wit R, Bouvier T. (2013). Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. *Environ Microbiol* **15**: 2463–2475.
McDaniel L, Paul JH. (2005). Effect of nutrient addition and environmental factors on prophage induction in natural populations of marine synechococcus species. *Appl Environ Microbiol* **71**: 842–850.
McGrath S, Fitzgerald GF, van Sinderen D. (2002). Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Mol Microbiol* **43**: 509–520.
McNair K, Bailey BA, Edwards RA. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**: 614–618.
Menouni R, Champ S, Espinosa L, Boudvillain M, Ansaldi M. (2013). Transcription termination controls prophage maintenance in Escherichia coli genomes. *Proc Natl Acad Sci USA* **110**: 14414–14419.
Middelboe M. (2000). Bacterial growth rate and marine virus-host dynamics. *Microb Ecol* **40**: 114–124.
Modi SR, Lee HH, Spina CS, Collins JJ. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**: 219–222.
Oliveira PH, Touchon M, Rocha EPC. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res* **42**: 10618–10631.
Paradis E, Claude J, Strimmer K. (2004). APE: analyses of phylogenetics and evolution in R596 language. *Bioinformatics* **20**: 289–290.
Pirofski LA, Casadevall A. (2012). Q and A: What is a pathogen? A question that begs the point. *BMC Biol* **10**: 6.
Pradeep Ram AS, Sime-Ngando T. (2010). Resources drive trade-off between viral lifestyles in the plankton: evidence from freshwater microbial microcosms. *Environ Microbiol* **12**: 467–479.
Ptashne M. (1992). *Genetic Switch: Phage Lambda and Higher Organisms*, 2nd edn, Blackwell: Cambridge, MA.
Roux S, Hallam SJ, Woyke T, Sullivan MB. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4**: 1–20.
Shan J, Korbsrisate S, Withatanung P, Adler NL, Clokie MR, Galyov EE. (2014). Temperature dependent bacteriophages of a tropical bacterial pathogen. *Front Microbiol* **5**: 599.
Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. (2010). Mobility of plasmids. *Microbiol Mol Biol Rev* **74**: 434–452.
Spearman C. (1904). The proof and measurement of association between two things. *Am J Psychol* **15**: 72–101.
St-Pierre F, Endy D. (2008). Determination of cell fate selection during phage lambda infection. *Proc Natl Acad Sci USA* **105**: 20705–20710.
Stearns SC. (1977). The evolution of life history traits: a critique of the theory and a review of the data. *Annu Rev Ecol Syst* **8**: 145–171.
Stewart FM, Levin BR. (1984). The population biology of bacterial viruses: why be temperate? *Theor Popul Biol* **26**: 93–117.
Torrella F, Morita RY. (1981). Microcultural study of bacterial size changes and microcolony and

ultramicrocolony formation by heterotrophic bacteria in seawater. *Appl Environ Microbiol* **41**: 518–527.
Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C. (2011). CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* **193**: 2460–2467.
Touchon M, Rocha EP. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* **24**: 969–981.
Touchon M, Rocha EP. (2010). The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. *PLoS ONE* **5**: e11126.
Vieira-Silva S, Rocha EPC. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* **6**: e1000808.
Vieira-Silva S, Touchon M, Abby SS, Rocha EP. (2011). Investment in rapid growth shapes the evolutionary rates of essential proteins. *Proc Natl Acad Sci USA* **108**: 20030–20035.
Volkmer B, Heinemann M. (2011). Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS One* **6**: e23126.
Wagner PL, Waldor MK. (2002). Bacteriophage control of bacterial virulence. *Infect Immun* **70**: 3985–3993.
Waldor MK, Friedman DI. (2005). Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol* **8**: 459–465.
Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM *et al.* (2010). Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* **1**: 147.
Weinbauer MG. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127–181.
Westra ER, Buckling A, Fineran PC. (2014). CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* **12**: 317–326.
Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics Bull* **1**: 80–83.
Wilhelm SW, Suttle CA. (1999). Viruses and nutrient cycles in the sea viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**: 781–788.
Williamson SJ, Houchin LA, McDaniel L, Paul JH. (2002). Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Appl Environ Microbiol* **68**: 4307–4314.
Winter C, Bouvier T, Weinbauer MG, Thingstad TF. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol Mol Biol Rev* **74**: 42–57.
Zeldovich KB, Berezovsky IN, Shakhnovich EI. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* **3**: e5.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)

Annexe 3 : Article 2 as contributing author

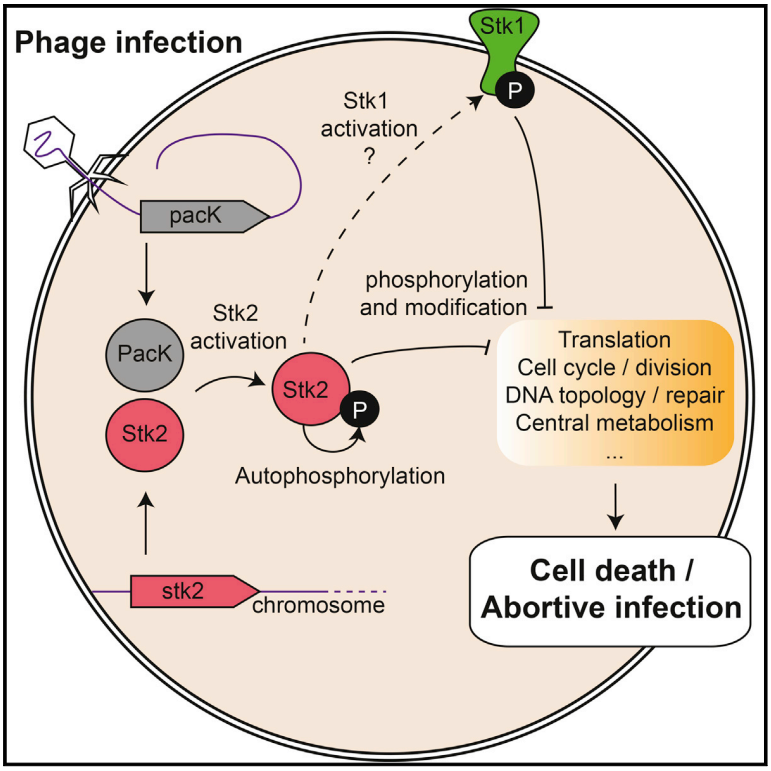
The following article describes the discovery of a new defense pathway in *Staphylococci*. The defense pathways relies on the activation of Stk2, a serine/threonine kinase by a phage protein called PacK. I contributed to Figure 1, supplementary Table S2 and supplementary Table S3 by describing the distribution of Stk2 in different *Staphylococci* and by identifying PacK homologs in diverse phages.

Article

Cell Host & Microbe

A Eukaryotic-like Serine/Threonine Kinase Protects Staphylococci against Phages

Graphical Abstract



Highlights

- Stk2 is a serine/threonine kinase involved in phage defense in Staphylococci
- Stk2 is activated by the PacK phage protein
- Activation of Stk2 leads to cell death, blocking phage propagation
- Stk2 mode of action is similar to some eukaryotic viral defense pathways

Authors

Florence Depardieu, Jean-Philippe Didier, Aude Bernheim, Andrew Sherlock, Henrik Molina, Bertrand Duclos, David Bikard

Correspondence

bertrand.duclos@univ-lyon1.fr (B.D.), david.bikard@pasteur.fr (D.B.)

In Brief

Serine/threonine kinases are critical for eukaryotic antiviral responses. Depardieu et al. now report a eukaryotic-like serine/threonine kinase in Staphylococci that protects against bacteriophages by triggering the death of infected cells. This abortive infection system is activated by a phage protein and leads to the extensive phosphorylation of essential cellular pathways.

Accession Numbers

KU598975

Depardieu et al., 2016, Cell Host & Microbe *20*, 471–481
October 12, 2016
http://dx.doi.org/10.1016/j.chom.2016.08.010

A Eukaryotic-like Serine/Threonine Kinase Protects Staphylococci against Phages

Florence Depardieu,[1] Jean-Philippe Didier,[2,6] Aude Bernheim,[1,3] Andrew Sherlock,[4] Henrik Molina,[5] Bertrand Duclos,[2,*] and David Bikard[1,7,*]
[1]Synthetic Biology Group, Microbiology Department, Institut Pasteur, Paris 75015, France
[2]Institute of Molecular and Supramolecular Chemistry and Biochemistry, University of Lyon-CNRS, Villeurbanne 69100, France
[3]AgroParisTech, Paris 75005, France
[4]Laboratory of Bacteriology
[5]Proteomics Resource Center
Rockefeller University, New York, NY 10065, USA
[6]Present address: Unilabs SA, 1296 Coppet, Switzerland
[7]Lead Contact
*Correspondence: bertrand.duclos@univ-lyon1.fr (B.D.), david.bikard@pasteur.fr (D.B.)
http://dx.doi.org/10.1016/j.chom.2016.08.010

SUMMARY

Organisms from all domains of life are infected by viruses. In eukaryotes, serine/threonine kinases play a central role in antiviral response. Bacteria, however, are not commonly known to use protein phosphorylation as part of their defense against phages. Here we identify Stk2, a staphylococcal serine/threonine kinase that provides efficient immunity against bacteriophages by inducing abortive infection. A phage protein of unknown function activates the Stk2 kinase. This leads to the Stk2-dependent phosphorylation of several proteins involved in translation, global transcription control, cell-cycle control, stress response, DNA topology, DNA repair, and central metabolism. Bacterial host cells die as a consequence of Stk2 activation, thereby preventing propagation of the phage to the rest of the bacterial population. Our work shows that mechanisms of viral defense that rely on protein phosphorylation constitute a conserved antiviral strategy across multiple domains of life.

INTRODUCTION

The arms race between bacteria and phages has led to the evolution of many bacterial defense systems that can act at every stage of the phage life cycle, blocking phage adsorption, DNA injection, degrading phage DNA, and interfering with phage replication or the production of phage proteins (Labrie et al., 2010). These defense systems are mechanistically diverse and can vary considerably among bacterial species or even among different isolates of a particular species. At a glance, bacterial defense against phages has little in common with eukaryotic antiviral systems. In plants, defense is primarily conducted via RNAi, while in vertebrates, pattern-recognition receptors (PRRs) recognize nucleic acids and proteins from pathogens and activate the interferon, proinflamatory, and adaptive immune responses (Kanneganti, 2010; Palm and Medzhitov, 2009; Pumplin and Voinnet, 2013; Sadler and Williams, 2008). Serine/threonine kinases (STKs) play a critical role at different stages of the antiviral response in both plants and vertebrates. They behave as switches that are activated by phosphorylation of one or several residues in an activation loop (Huse and Kuriyan, 2002). Some STKs, such as the interferon-induced, dsRNA (double-stranded RNA)-activated protein kinase (PKR) in mammals, can also directly sense and interfere with viruses (Yan and Chen, 2012). Upon activation by dsRNA, PKR phosphorylates the translation initiation factor eIF2α, blocking translation and viral protein synthesis. A similar mechanism was also recently described in plants. The NIK1 STK of *Arabidopsis* was shown to phosphorylate the ribosomal protein L10 and globally suppress translation as an antiviral immunity strategy (Zorzatto et al., 2015).

STKs were assumed for a long time to exist only in eukaryotes, but eukaryotic-like STKs (eSTKs) have now been found in most bacterial clades, where they have been implicated in a variety of functions including cell-cycle control, exit of dormancy, cell wall synthesis, cell division, control of the central metabolism, and virulence (Pereira et al., 2011). Unlike eukaryotes, bacteria are generally not known to use STKs in viral defense. An exception to this is the *pgl* phage defense system from *Streptomyces coelicolor* and the related BREX (bacteriophage exclusion) systems, which are thought to exist in many unrelated bacteria (Goldfarb et al., 2015; Hoskisson et al., 2015; Sumby and Smith, 2002). These systems include an STK known as PglW for which kinase activity was confirmed in vitro; however, its exact role in the defense pathway is not yet understood (Hoskisson et al., 2015). Also of note is the discovery of a prophage-encoded tyrosine kinase that excludes superinfection by phage HK97 in *Escherichia coli* (Friedman et al., 2011).

Here we report the discovery of an eSTK involved in phage defense in *Staphylococci*. The Stk2 protein is activated when a specific phage protein, PacK, is present in the cell. The activation of Stk2 results in cell death through phosphorylation of proteins involved in essential cellular processes, including translation, transcription, control of cell cycle, and others. Infected cells

die before releasing infectious phages, thereby protecting neighboring bacteria. This altruistic defense strategy is known as abortive infection (Abi) and can be performed by mechanically diverse systems (Chopin et al., 2005). *Staphylococci* carry another STK known as Stk1 or PknB, which is important for cell wall structure, antimicrobial resistance, and virulence (Beltramini et al., 2009; Débarbouillé et al., 2009; Donat et al., 2009; Truong-Bolduc et al., 2008). The existence of a second STK, known as Stk2, present in only some isolates of *S. aureus*, was noted in a few studies, but its function remained mysterious (Débarbouillé et al., 2009; Didier et al., 2010). We now demonstrate that Stk2 provides immunity against bacteriophages through an Abi mechanism. Interestingly, the Stk1 kinase is also involved in this defense pathway, suggesting a phosphorylation cascade reminiscent of eukaryotic viral defense pathways.

RESULTS

Discovery of *stk2*, a Bacteriophage Defense Kinase

We isolated a temperate phage of *Staphylococcus epidermidis*, CNPx, a 43 kb Siphoviridae with 90.2% overall nucleotide identity to phage CNPH82 (Daniel et al., 2007). CNPx was isolated on strain LM1680, a derivative of *S. epidermidis* RP62A carrying a large deletion that includes a type III-A CRISPR system and a type I restriction modification (RM) system (Hatoum-Aslan et al., 2014). Interestingly, CNPx does not infect strain RP62A, suggesting that the region deleted in LM1680 contains a defense system, possibly the CRISPR or the type I RM, that blocks infection by this phage. To narrow down the position of the defense system, we tested the ability of CNPx to infect a collection of RP62A derivatives, obtained by Marraffini and colleagues, that carry various deletions of this region (Jiang et al., 2013). This analysis allowed us to identify an ~16 kb candidate region that carries the phage defense system (Figure 1A). To our surprise, this region did not include the type I RM system or the CRISPR system; instead, it contained a number of hypothetical proteins as well as an operon involved in potassium transport and associated regulatory genes (Table S1, available online). Genes coding for the hypothetical proteins were cloned either alone or two at a time on plasmid pC194, which is present in ~15 copies in the cell (Novick, 1989). The resulting plasmids (pDB31, pDB32, pDB33, and pDB34) were introduced by electrotransformation in strain LM1680, and the bacteria were tested for sensitivity to phage CNPx. The pDB31 plasmid-carrying gene SERP2479 provided strong resistance (efficiency of plaquing [EOP] < 10^{-5}), while the other genes carried by plasmids pDB32, pDB33, and pDB34 did not have any effect on the susceptibility of the strain to the phage (Figure 1B).

SERP2479 contains an STK domain that is easily identified by a prediction algorithm such as hhmer or CD-search (Figure 2A) (Finn et al., 2011; Marchler-Bauer and Bryant, 2004). Proteins identical to SERP2479 are found in several strains of *Staphylococcus aureus* (Figure 1C). In particular, protein SA0077, whose sequence is 100% identical to SERP2479, was previously described in *S. aureus* strain N315. It was named Stk2 and shown to phosphorylate the virulence regulator SarA in vitro, but could not be assigned a physiological role (Didier et al., 2010). Closely related proteins are also present in more distantly related Firmicutes, including some *Bacilli* and *Streptococci* (Table S2). In addition to the kinase domain, a distinctive feature of these proteins is the presence of large N-terminal and C-terminal domains of unidentified fold or function.

Stk2 Provides Phage Resistance in Staphylococci

To investigate the scope and significance of our finding, we decided to test whether *stk2* could also block phage infection in *S. aureus*. The pDB31 plasmid, carrying *stk2* under the control of its wild-type promoter, was introduced by electrotransformation in several *S. aureus* strains (TB4, NCTC8325-4, and RN4220) that do not carry *stk2* in their chromosome (Bae et al., 2006; Nair et al., 2011). Stk2 provided strong resistance to phage phiNM1 in all backgrounds (Table S3). Strain RN4220 is a derivative of *S. aureus* strain NCTC8325, which is easy to manipulate but is generally not considered to be a good model strain for the study of bacterial virulence (Nair et al., 2011). Nonetheless, since the phage resistance phenotype provided by *stk2* was identical in all tested strains, we decided to use RN4220 for the remainder of this study. We tested the ability of plasmid pDB31 to confer resistance to infection by different phages in the RN4220 background, including five Siphoviridae (phage 80alpha, phage 85, phiNM1, phiNM2, and phiNM4) and one Twort-like Myoviridae (phage Staph1N) (Bae et al., 2006; Łobocka et al., 2012). Stk2 provided resistance against all tested Siphoviridae, but not against the Twort-like phage (Table S3).

Characterization of the Stk2 Kinase Activity

The *stk2* gene was cloned with a 6x His N-terminal tag in plasmid pET15b and introduced in *E. coli* BL21 (DE3). Upon induction with IPTG, efficient overproduction of His6-Stk2 fusion protein was obtained, though in the form of inclusion bodies. The His6-Stk2 product was then purified by a denaturation/renaturation method using guanidinium chloride, followed by a step of purification on an affinity column. Finally, the linked His6 was removed through proteolysis by thrombin (Figure 2B). Autophosphorylation activity was tested in the presence of various divalent cations: Mn2+, Mg2+, Ca^{2+}, Zn2+, and Co2+ (Figure 2C). It was observed that purified Stk2 was significantly labeled in vitro in the presence of [γ-32P] ATP and Mn2+ (Figure 2C, lane 4). The ability of Stk2 to autophosphorylate in these conditions indicates that it displays intrinsic kinase activity. To exclude the possibility of contamination by an exogenous kinase, the invariant lysine 152, involved in the binding of the ATP phosphoryl donor (Figure 2A), was mutated to isoleucine. As expected, Stk2-K152I could no longer autophosphorylate (Figure 2B, lane 5).

The phosphoamino acid content of the labeled protein was determined after acid hydrolysis and two-dimensional analysis (Duclos et al., 1991). Both phosphoserine and phosphothreonine were revealed on the corresponding autoradiogram (Figure 2D), indicating that Stk2 was modified exclusively on these two types of residues. NanoLC/nano-spray/tandem mass spectrometry was then used for the identification of phosphorylated peptides and for the localization of the phosphorylation sites in Stk2 (Molle et al., 2006). Nine phosphorylation sites could be identified, including three sites in the activation loop of Stk2 (S272, T275, and T278) (Figure 2A). Various mutated proteins

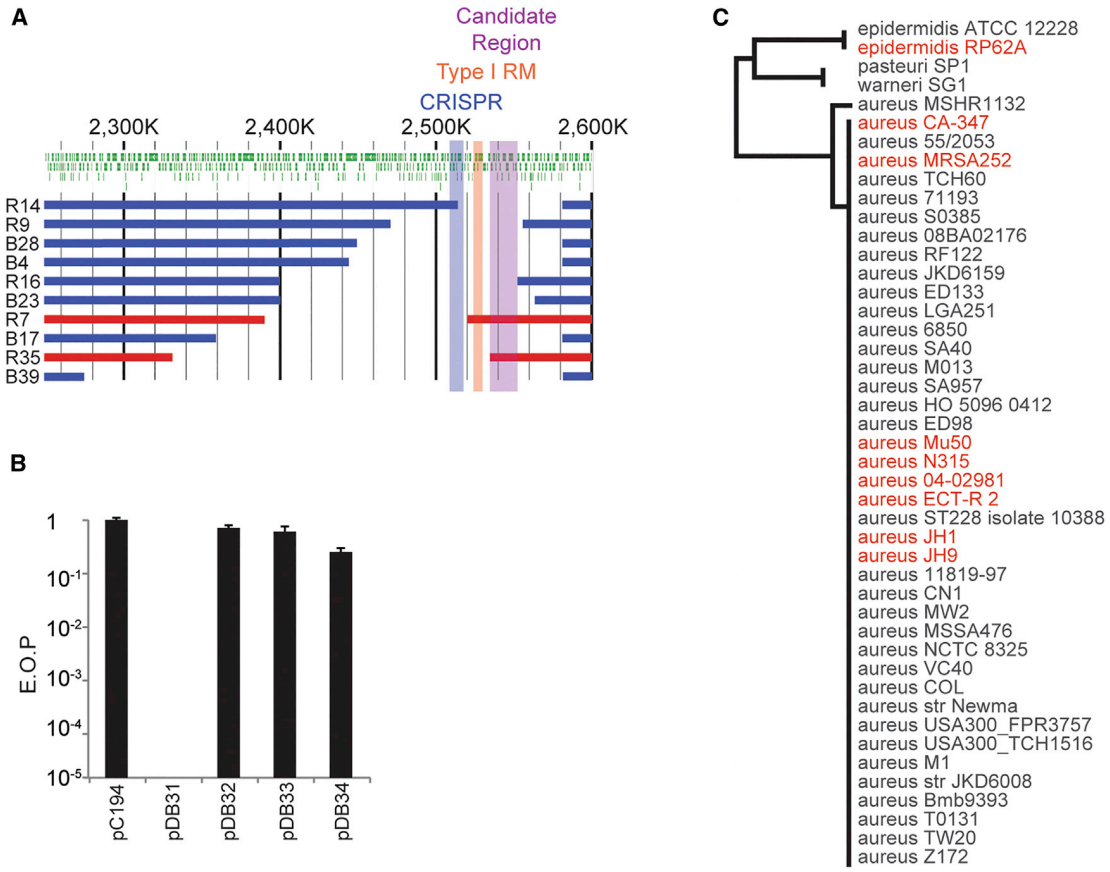


Figure 1. Discovery of a Phage Defense System in *S. epidermidis* Strain RP62A

(A) Strains with various deletions in the region of the CRISPR locus (numbered as in Jiang et al., 2013) were screened for sensitivity to bacteriophage CNPx. Open reading frames (ORFs) in the genomic region are represented in green. Each horizontal line represents a strain and the line is discontinued in the deleted region. Blue lines indicate strains sensitive to phage CNPx while red lines indicate resistance. The region that is sufficient to provide resistance is highlighted in purple. The exact position of the candidate region is 2535598–2551561 (GenBank: NC_002976) and the ORFs it contains are described in Table S1.

(B) ORFs contained in the candidate region were cloned either alone or two at a time on plasmid pC194 to give plasmids pDB31 (*serp2479*, *stk2*), pDB32 (*serp2480*), pDB33 (*serp2481* + *serp2482*), and pDB34 (*serp2483* + *serp2484*) (see Table S1). Efficiency of plaquing (EOP) of phage CNPx is reported on *S. epidermidis* strain LM1680 containing these different plasmids (mean + SD, n = 3). No plaques were recovered in bacteria carrying gene *Serp2479* (*stk2*) (EOP detection limit of 10^{-5}).

(C) A tree of Staphylococci was constructed based on the assembled complete genomes available in GenBank. Red indicates the presence of proteins with a minimum of 90% identity to Stk2 (SERP2479) over 100% of the sequence length. More distantly related proteins were also identified in Staphylococci that are not present in this tree and are reported in Table S2.

were produced and purified, and the effect of substitution of the different residues to alanine on kinase activity was analyzed by measuring autophosphorylation activity (Figure 2E). Kinase activity was completely abolished when the substitution was on T147, S272, T275, and T278. In contrast, substitution of S246 had no effect on the activity of Stk2, and replacement of S163, S337, and S360 unexpectedly increased Stk2 activity. Altogether, these results show that Stk2 is indeed an active STK.

To confirm in vivo that the kinase activity of Stk2 is required for the phage defense phenotype, we mutated conserved residues: K152 in the ATP-binding loop, the predicted catalytic aspartate D243, and the T275 autophosphorylation residue in the activation loop (Figure 2A). All alleles were introduced in *S. aureus* strain RN4220 and tested for sensitivity to phage phiNM1 (Figure 3A). As expected, all mutants showed sensitivity to the phage.

Stk2 Triggers Cell Death

Several assays were performed to understand the mechanism of protection provided by Stk2. We first tested whether it could affect phage adsorption. Cells carrying plasmid pDB31 or the control pC194 were both able to adsorb 99% of the phiNM1 phage particles. However, infection of growing cells carrying *stk2* (pDB31) by phage phiNM1 at high MOI led to an interruption in the growth of the culture (Figure 3B). This result stands in clear contrast to the lysis observed for cells that lack *stk2* and are sensitive to the phage. Consistent with this observation, ~3,000× fewer phage particles are recovered after infection of cells carrying *stk2* relative to cells without *stk2* (Figure 3C). Also, measurement of the efficiency of center of infection (ECOI) in the presence of *stk2* indicates that only 0.4% ± 0.08% of cells receiving the phiNM1 phage are able to release functional phiNM1 particles. Plating a culture of cells carrying *stk2* after infection reveals that most cells are dead; only ~5% of cells

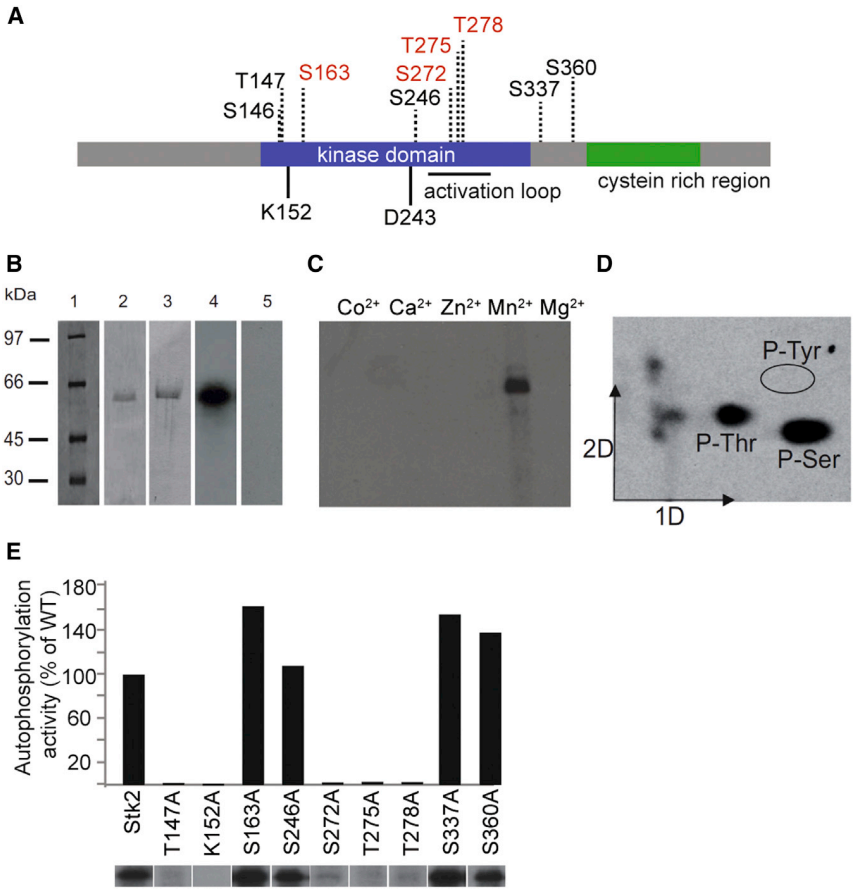


Figure 2. Stk2 Is an Active Serine/Threonine Kinase

(A) Schematic presentation of the position of phosphoresidues in protein Stk2 and location of different domains. The kinase domain, the activation loop, the cystein-rich region, and the position of the residues K152 and D243 are shown. The position of phosphoresidues is indicated by dotted bars. Residues in red were also shown to be phosphorylated in vivo (see Table S7).

(B) SDS-PAGE analysis of purified Stk2 (lane 2) and Stk2 mutant K152I (lane 3) after staining with Coomassie blue. Molecular mass standards are shown on the left (lane 1). Autophosphorylation of Stk2 (lane 4) and mutant K152I (lane 5) in the presence of radioactive [γ-32P] ATP is shown. Radioactive molecules were detected by autoradiography.

(C) Effect of cations on Stk2 autophosphorylation activity in vitro.

(D) Two-dimensional analysis of phosphorylated amino acids in Stk2. The acid-stable phosphoamino acids from [γ-32P]-labeled Stk2 were separated by electrophoresis in the first dimension (1D), followed by ascending chromatography in the second dimension (2D). P-Tyr (phosphotyrosine), P-Ser (phosphoserine), and P-Thr (phosphothreonine) were located by ninhydrin staining. Phosphorylated molecules were revealed by autoradiography.

(E) Effect of mutations on the kinase activity of Stk2. Purified wild-type and mutants of Stk2 were incubated in the presence of radioactive [γ-32P] ATP, proteins were separated by SDS-PAGE, and radioactive molecules were detected by autoradiography.

form colonies (Figure 3C). This shows that *stk2* mediates cell death and acts as an Abi system, killing bacteria upon infection to prevent phage propagation (Abedon, 2012).

It is worth nothing that the number of bacteria that survive phiNM1 infection is similar in the presence and absence of *stk2* (Figure 3C). Wild-type RN4220 cells can survive phiNM1 infection when the phage enters lysogeny and integrates in the genome. Interestingly, analysis of cells that survive phiNM1 infection in the presence of *stk2* revealed that some (3/8) had lysogenized phage phiNM1, while the remainder (5/8) most likely did not receive a phiNM1 phage particle, eliminated the phage without dying, or mutated its receptor (Figure S1A). These results suggest that *stk2* kills staphylococci only if the phage enters its lytic cycle. To confirm this, we sought to induce the lytic cycle of phage phiNM1 lysogenized in cells carrying *stk2* or a control plasmid (Figure 3D). Prophages were induced with mitomycin C. In the presence of *stk2*, the culture stopped growing but did not lyse. In agreement with this observation, $3*10^{4}$-fold fewer phage particles were recovered after induction of cells carrying *stk2* compared to cells carrying the control plasmid (Figure 3D). Note that in the absence of mitomycin C, phage phiNM1 is spontaneously induced at a lower rate. Under these conditions, the presence of *stk2* also limits the number of phages released. Altogether, these experiments demonstrate that presence of phage DNA is not recognized by Stk2; instead, the Abi phenotype of Stk2 is only triggered during the lytic cycle of the phage.

Identification of the Stk2 Activation Factor

In the pDB31 plasmid used here, *stk2* is expressed under the control of its wild-type promoter. To understand whether the transcriptional control of *stk2* is important for the Abi phenotype, we cloned *stk2* under the control of a Ptet promoter (Table S5). Resistance to phage was only observed upon induction of Stk2 with anhydrotetracycline (aTc), and overexpression of Stk2 on its own did not lead to any growth defect (Figure S1B). These results clearly show that the Abi phenotype is only induced in the presence of the phage, and that the natural transcriptional control of *stk2* is not required. Thus, as expected, Stk2 likely behaves as a protein switch that is activated upon phage infection. While Stk2 is able to rapidly autophosphorylate in vitro, we believe that it is not active in the absence of phage infection in vivo.

To understand what might activate Stk2, we isolated phage mutants capable of infecting *S. aureus* strain RN4220 carrying pDB31. The EOP of phage phiNM1 on cells carrying *stk2* is only $5*10^{-7}$ (Table S3). Nonetheless, some plaques can be recovered and propagated on cells expressing *stk2* (Figure S2A). Phages isolated in this way retain the ability to infect cells carrying *stk2* even after being passaged on cells lacking *stk2*, suggesting that the new phenotype is the result of mutation

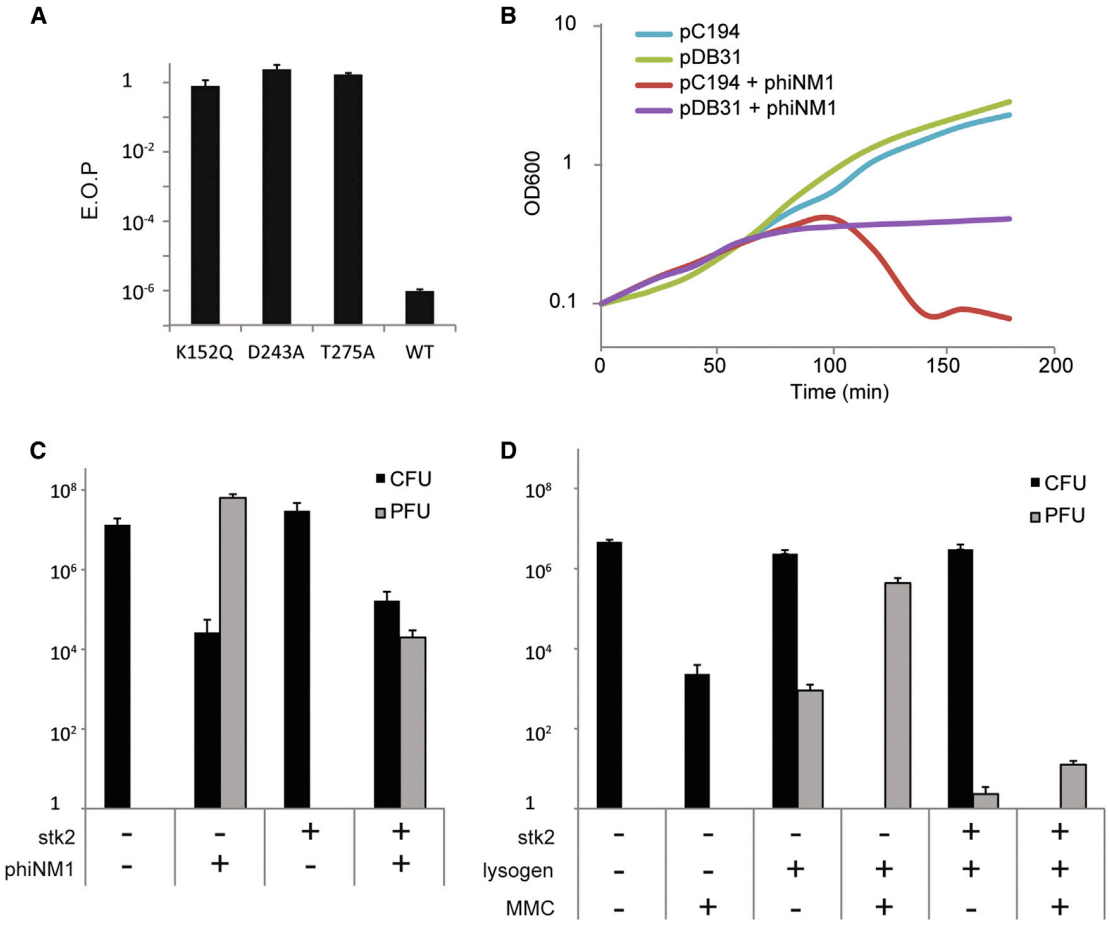


Figure 3. The Stk2 Kinase Triggers Cell Death during the Phage Lytic Cycle

(A) EOP of phage phiNM1 on RN4220 *S. aureus* cells in the presence of the wild-type Stk2 protein (pDB31) or various mutants: K152Q (pDB81), D243A (pDB82), and T275A (pDB83).

(B) Growth curve of *S. aureus* RN4220 carrying control plasmid pC194 or pDB31 (pC194Ω*stk2*). Phage phiNM1 was added after 1 hr of growth at an MOI of 10.

(C) PFUs and colony-forming units (CFUs) recovered after infection of RN4220 cells carrying *stk2* or not. Cells were grown to OD = 0.2 and incubated with phiNM1 for 2 hr. Cells were then plated on TSA to measure CFUs, and the filtered supernatant was spotted on a top agar lawn of RN4220 cells to measure PFUs. Upon infection with phiNM1 and in the presence of *stk2*, cells are killed, but the phage is not amplified.

(D) PFUs and CFUs recovered after induction with mitomycin C (MMC) of growing RN4220 carrying a phiNM1 lysogenic phage or not, in the presence or absence of *stk2*. Upon induction, cells carrying a lysogenic phage are killed regardless of the presence of *stk2*, but the production of phage is inhibited in the presence of *stk2*. Note that PFUs are recovered even in the absence of MMC due to the spontaneous induction of the phage. See also Figure S1.

Error bars represent the SD of three replicates.

and not epigenetic variation. These phiNM1 mutants form small plaques and are harder to propagate than the wild-type phage (data not shown). Two mutant phages were sequenced (phiNM1k1 and phiNM1k2), and both carried independent missense mutations (V200M and K30E) in the same gene, SAPPV1_GP14. These results indicate that this protein likely activates Stk2. SAPPV1_GP14 contains a P loop NTPase domain frequently found in proteins involved in molecular motion. An interesting observation enabled us to link the function of this gene to either DNA replication or packaging. When sequencing phage phiNM1k1 and phiNM1k2, we observed that coverage was highest shortly after the packaging site of the phage and then slowly dropped over the rest of the sequence (Figure 4A). This contrasts sharply with the wild-type phage, which shows almost uniform coverage throughout the sequence. A possible explanation for this skewed coverage is that the capsids from which DNA was purified do not all contain the full phage genome; instead, most capsids only contain the part of the phage genome that is packaged first. Phage particles that contain only part of the genome would likely not be functional and might even lack the tail. As we did not purify intact phage particles before DNA extraction, we recovered DNA from both functional phage particles and any incompletely assembled particles present in our samples. Thus, the skewed coverage could be explained by random premature termination of phage DNA packaging. Such premature termination could be due to a defect in the packaging machinery, but could also come from problems with concatemer formation or DNA replication. The position of SAPPV1_GP14

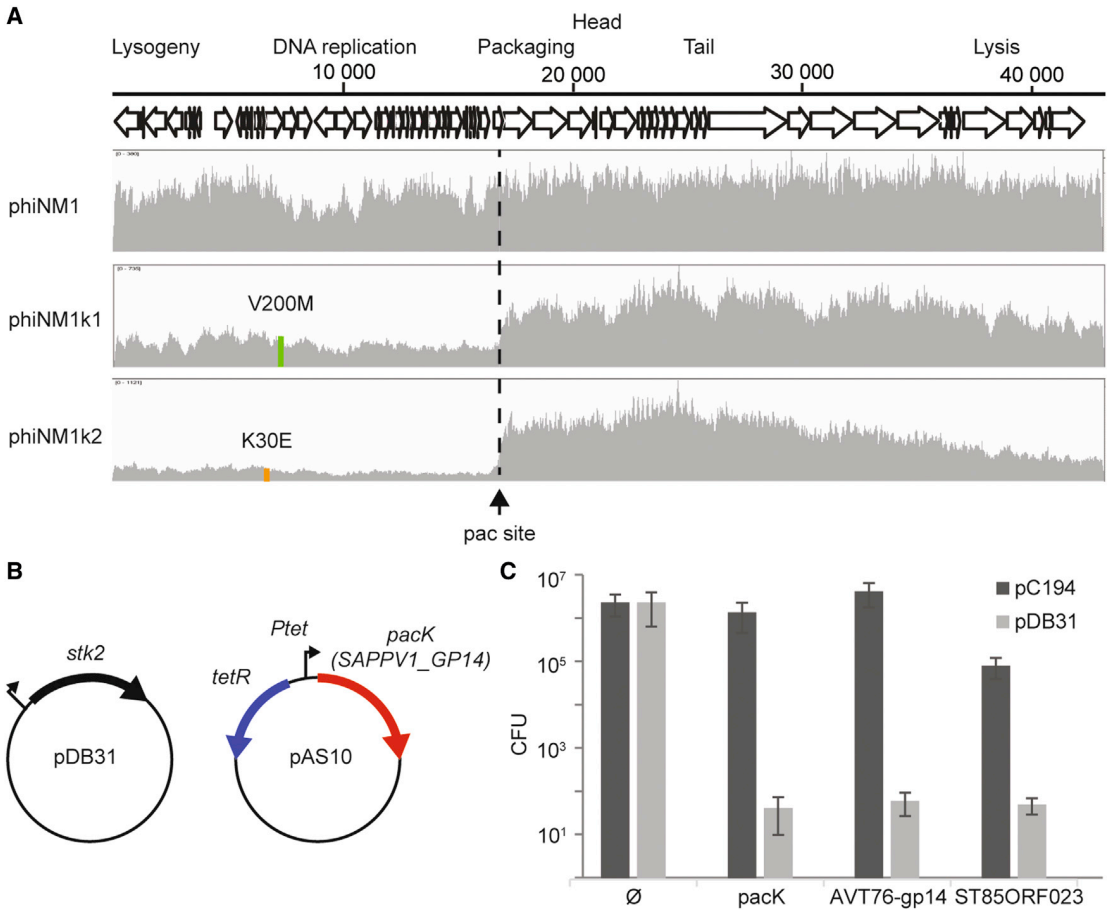


Figure 4. A Phage Protein Triggers Stk2-Mediated Cell Death

(A) Two mutants of phage phiNM1 able to infect RN4220 cells carrying *stk2* were isolated and sequenced. Coverage along the phage genome is plotted and mutations are highlighted. Both phages carry a mutation in gene *pacK* (SAPPV1_GP14) V200M and K30E for phages phiNM1K1 and phiNM1K2, respectively.

(B) To test whether PacK is sufficient to activate Stk2, the *pacK* gene was cloned under the control of a Ptet promoter on a pE194 vector giving pAS10, and introduced in cells carrying plasmid pDB31.

(C) RN4220 cells carrying gene *pacK* (SAPPV1_GP14) from phage phiNM1 (pAS10), AVT76_gp14 from phage phiNM2 (pFD16), or ST85ORF023 from phage 85 (pFD20) under the control of a Ptet promoter, together with plasmids pC194 or pDB31, were grown to OD ~ 0.2 and induced with aTc. After 1 hr of induction, cells were plated and colonies quantified (mean + SD, n = 3). See also Figure S2.

in the replication cluster of the phage supports the later hypothesis.

Because of its phenotype in DNA packaging, we decided to call the SAPPV1_GP14 gene *pacK*. To confirm that this phage protein is sufficient to trigger Stk2, we cloned *pacK* under the control of an inducible Ptet promoter on plasmid pE194, giving plasmid pAS10 (Figure 4B). Upon induction with aTc, cell death was observed only when *stk2* was present in the cells (Figure 4C). The V200M and K30E mutations identified in the mutant phages were also tested in this assay and abolished the Abi phenotype (Figure S2B). These experiments confirm that PacK is sufficient to trigger Stk2-mediated cell death.

It is interesting to note that Stk2 can provide resistance to phages that do not carry PacK (see Tables S3 and S4), suggesting that it can be activated by other phage proteins. In particular, this is true for *S. epidermidis* phage CNPx, which was used in this study to first identify Stk2 (Figure 1), as well as *S. aureus* phage 80alpha, phage 85, and phiNM2 (Table S3). In phage phiNM2, gene AVT76_GP14 encodes a protein with 43% identity to PacK, but no homologous proteins exist in phage 80alpha, phage 85, or CNPx. To identify how these phages activate Stk2, we isolated mutants of phiNM2 and phage 85 capable of infecting *S. aureus* cells expressing *stk2*. Sequencing of these mutants revealed an H230T mutation in gene AVT76_GP14 of phage phiNM2 and a K97G mutation in gene ST85ORF023 of phage 85. To confirm that these phage genes encode activators of Stk2, we cloned them under the control of a Ptet promoter on plasmid pE194, producing plasmids pFD16 and pFD20, respectively (Table S5). After induction with aTc, cells were killed in the presence of *stk2*, but not in its absence (Figure 4C). The mutations identified in these genes were also confirmed to abolish Stk2-mediated cell death (data not shown). These results show that, in addition to PacK, two other phage proteins can activate Stk2.

Identification of Stk2 Phosphorylation Targets

Our results suggest that Stk2 triggers an Abi phenotype through phosphorylation of one or several host proteins. To identify the phosphorylation target(s) of Stk2, we first characterized *S. aureus* colonies that survive the induction of *pacK* in the presence of Stk2, with the goal of identifying mutants of the target proteins. Unfortunately, all of the 36 analyzed colonies carried mutations either in the *stk2* or *pacK* genes, but no other mutant could be identified (data not shown). This result suggests that several mutations might be required to survive Stk2 activation; these would occur at a lower frequency than point mutations in *stk2* or *pacK*.

We then performed a phosphoproteome analysis of cells expressing *pacK* in either the presence or absence of *stk2*. Expression of PacK was induced from plasmid pAS10 (Ptet-*pacK*) in exponentially growing cells. After 30 min of induction, proteins were precipitated and digested followed by titaniumdioxide-based phosphopeptide enrichment (Larsen et al., 2005). To confidently differentiate basal and Stk2-induced phosphorylation events, we labeled the different proteomes with mass spectrometry-differentiable stable isotopes of dimethyl (Boersema et al., 2009). We identified 32 phosphopeptides that could only be found in the presence of Stk2 (Table 1). These include several proteins related to translation, including elongation factors Tu and P; 50S ribosomal proteins L6, L5, and L31; and the MetG methionine-tRNA ligase. This extensive phosphorylation of the translation machinery likely indicates that translation is modified after Stk2 activation. In addition, we identified proteins involved in global transcription control, cell-cycle control, stress response, DNA topology, DNA repair, and central metabolism. This suggests a coordinated response influencing many aspects of the cellular machinery, and a general shift toward stress response and growth arrest. Phosphopeptides corresponding to Stk2 itself could also be identified. Three residues are phosphorylated in the activation loop (S272, T275, and T278), as well as a serine S163 between the P loop and the catalytic site. These residues were also identified in the in vitro autophosphorylation assay (Figure 2). Phosphorylation of the trigger PacK protein could also be identified at residue S176, suggesting that Stk2 interacts with PacK directly. Mutation of the S176 residue to alanine did not have any impact on the Abi phenotype (Figure S3).

Role of Stk1 in Stk2-Mediated Immunity

The question of whether the phosphopeptides identified are directly phosphorylated by Stk2 remains to be investigated. Indeed, it is possible that the activation of Stk2 results in the activation of Stk1, which would lead to secondary phosphorylation events. For instance, the elongation factor P identified in our analysis was previously reported as a target of Stk1 (Lomas-Lopez et al., 2007). In further support of this possible role of Stk1 in the phage defense phenotype, we detected the phosphorylation of Stk1 at two threonines in the activation loop (T164 and T166). The phosphopeptide carrying these residues was 1.5-fold more abundant in the presence of Stk2 than in its absence (Table S7). To investigate whether Stk1 could play a role in Stk2-mediated phage defense, plasmid pDB31 carrying *stk2* was introduced by electrotransformation in *S. aureus* strain NCTC8325-4 and in the corresponding *stk1* deletion mutant (Débarbouillé et al., 2009). The EOP of phage phiNM1 on cells carrying both *stk1* and *stk2* is $4.4*10^{-6}$, but when only *stk2* is present, the EOP jumps to $2*10^{-2}$ (Figure 5A). Thus, in the absence of *stk1*, we can still observe some protective effect of *stk2*, but ultimately *stk1* is required for efficient immunity. We also investigated whether *stk2* could trigger cell death in the absence of *stk1*. *S. aureus* NCTC8325-4 and the *stk1* mutant were electrotransformed with both pDB31 (*stk2*) and pAS10 (Ptet-*pacK*) plasmids. Upon induction of PacK expression, cells were killed with identical efficiencies regardless of the presence of *stk1* (Figure 5B). This demonstrates that while Stk1 is necessary for efficient antiviral immunity, it is not required for Stk2-mediated cell death.

DISCUSSION

Recent bioinformatics analyses have led to the discovery that bacteriophage defense systems frequently cluster together in bacterial genomes (Makarova et al., 2011). Here we report the discovery of a defense system in close proximity to the type III CRISPR and type I RM system of *S. epidermidis* RP62A. SERP2479, or Stk2, is responsible for Abi and cell death upon phage infection. Stk2 belongs to the family of eukaryotic-like STKs but differs from previously described eSTKs in its lack of transmembrane or PASTA domains. We were able to confirm the kinase activity of Stk2 in vitro and identified nine autophosphorylated residues. Four of these residues were corroborated in vivo, including three in the activation loop (S272, T275, and T278), as well as a serine (S163) close to the ATP-binding region (Figure 2A). All three residues of the activation loop are essential for in vitro autophosphorylation, while an S163A mutation actually increased the kinase activity, suggesting a regulatory role. The phosphorylation of several residues in the activation loop of eSTKs has been reported before and seems to be a common feature of these kinases (Young et al., 2003).

Our data suggest that while Stk2 is able to autophosphorylate in vitro, it is only activated in the presence of a phage protein in vivo. We found three such phage proteins by analyzing the genomes of mutant phages able to propagate on cells carrying *stk2*: gene SAPPV1_GP14 (*pacK*) from phage phiNM1; gene AVT76_GP14 from phage phiNM2, a distant homolog of *pacK* with 43% protein identity; and gene ST85ORF023 from phage 85, which shows no identity to PacK. The PacK protein carries a P loop NTPase domain and leads to a defect in phage DNA packaging when mutated. This defect could either be due to problems in DNA replication leading to DNA molecules in a bad conformation for packaging, or to a defect in packaging itself. It is currently identified in databases as the chromosomal replication initiator DnaA. However, we believe this to be a simple case of incorrect annotation, as no significant homologies can be found between PacK and DnaA proteins. The function of ST85ORF023 is not known, and no protein domain of known function can be identified. These activator genes are located within the phage lytic operon, which likely explains why Stk2-induced cell death is not triggered when the phage enters lysogeny (Figure 3D). However, the induction of a lysogenic phage in cells carrying *stk2* also leads to cell death. The ability of Stk2 to tolerate prophages while maintaining an active defense against the phage lytic cycle is reminiscent of the similar

Table 1. List of Proteins Phosphorylated upon Expression of PacK and in the Presence of Stk2 that Are Never Found Phosphorylated in the Absence of Stk2

	Gene Name	Protein Description	Accession Number
Transcription	greA	transcription elongation factor GreA	UniProt: A6QHF1
	sigA	RNA polymerase sigma factor SigA	UniProt: P0A0J0
	nusA	transcription termination/antitermination protein NusA	UniProt: Q2G2D2
Translation	metG	methionine-tRNA ligase	UniProt: A6QEE3
	tuf	elongation factor Tu	UniProt: A6QEK0
	efp	elongation factor P	UniProt: A6QH73
	rpmE2	50S ribosomal protein L31 type B	UniProt: A6QIW4
	rplF	50S ribosomal protein L6	UniProt: A6QJ77
	rplE	50S ribosomal protein L5	UniProt: A6QJ80
Cell cycle	ftsZ	cell division protein FtsZ	UniProt: A6QG86
	sepF	cell division protein SepF	UniProt: Q2FZ86
	gpsB	cell-cycle protein GpsB	UniProt: Q2FYI5
Stress response	clpX	ATP-dependent Clp protease ATP-binding subunit ClpX	UniProt: Q2FXQ7
	clpB	chaperone protein ClpB	UniProt: Q2FZS8
	ydaG/yzzA	general stress protein 26	UniProt: Q2FVN7
	AQ00_RS06590	alkaline shock protein (Asp23)/stress response regulator gls24 homolog	UniProt: Q2FZ59
	telA	tellurite/ toxin anion resistance protein	UniProt: Q2FYM7
DNA topology and repair	ssb	single-stranded DNA-binding protein	UniProt: A6QE48
	parE	DNA topoisomerase 4 subunit B	UniProt: A6QGQ7
	mutS	DNA mismatch repair protein MutS	UniProt: Q2FYZ9
Central metabolism and biosynthesis	hemL1	glutamate-1-semialdehyde 2,1-aminomutase 1 (protoporphyrin-IX biosynthesis)	UniProt: A6QHK1
	pgk	phosphoglycerate kinase (glycolysis)	UniProt: A6QF82
	pgi	glucose-6-phosphate isomerase (glycolysis pathway)	UniProt: Q2FZU0
	dltA	D-alanine–poly(phosphoribitol) ligase subunit 1 (LTA biosynthesis)	UniProt: Q2FZW6
Others	pacK / SAPPV1_GP14	pacK (Stk2 activation protein)	UniProt: A6QDW1
	stk2	SERP2479	UniProt: Q5HK71
	ylaL	uncharacterized protein	UniProt: A6QFW6
	esxA	virulence factor EsxA (ESAT-6-like protein)	UniProt: Q2G189
	obg	GTPase	UniProt: Q2FXT1
	phoP	alkaline phosphatase synthesis two-component response regulator	UniProt: Q2FXN6
	AQ00_RS00105	short-chain dehydrogenase	UniProt: Q2FV41

A detailed list of all phosphopeptides identified with match metrics and measured ratios is provided in Table S7.

capacity of type III CRISPR systems to tolerate lysogenic phages (Goldberg et al., 2014).

An analysis of Staphylococcal phages in the RefSeq database shows that 23% carry homologs of PacK, while 13% carry homologs of ST85ORF023 (Table S4). Phage CNPx, which was used in this study to first identify Stk2 in *S. epidermidis*, carries a protein with 85% identity to ST85ORF23. Interestingly, a blast analysis only identified activators of Stk2 in Siphoviridae; none were found in other phage families. The ability of Stk2 to sense multiple phage proteins is fascinating and leads to deeper questions about the regulation of the kinase activity.

We hypothesize that the activation of Stk2 by PacK is the result of a direct interaction between the two proteins, as PacK is itself phosphorylated during the response. After sensing the presence of the infecting phage, the activation of Stk2 results in the phosphorylation of a large number of proteins involved in several core functions of the cell, including translation, transcription, and cell division (Table 1). The modulation of the activity of these proteins through phosphorylation is likely responsible for cell death, preventing phage propagation concurrently. It is worthwhile to note that while the SarA protein was previously reported as a phosphorylation target of Stk2, we did not identify it in our analysis (Didier et al., 2010). This could be explained by the fact that a serine-rich segment of SarA was not covered by our mass spectrometry analysis (data not shown).

Another STK known as Stk1 (sometimes also named PknB or PrkC), present in all Staphylococci, is also involved in this antiviral defense pathway. Deletion of *stk1* strongly impairs the ability of Stk2 to protect *S. aureus* against phages (Figure 5A). Nonetheless, the activation of Stk2 leads to cell death even in the absence of Stk1 (Figure 5B). Thus, the role of Stk1 in this defense pathway is likely to ensure that phage particles are not produced before

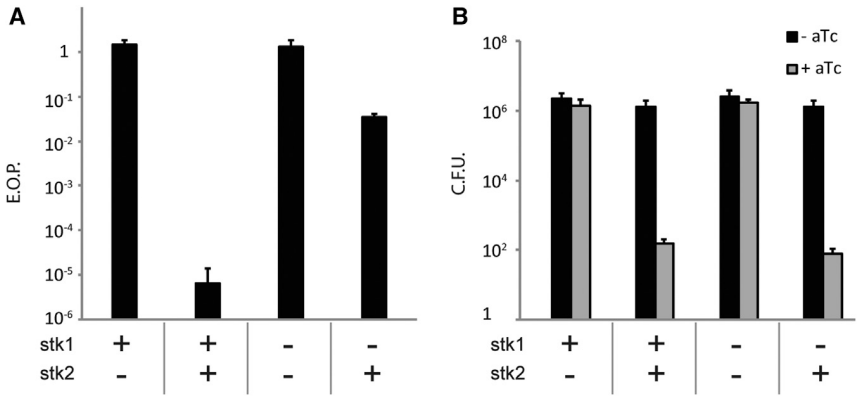


Figure 5. Stk1 Is Required for Efficient Stk2-Mediated Immunity

(A) EOP of phage phiNM1 is reported against *S. aureus* strain NCTC8325-4 harboring *stk1* or not (Δ*stk1*) in the presence of *stk2* (pDB31) or a control plasmid (pC194) (mean + SD, n = 3).

(B) NCTC8325-4 cells harboring *stk1* or not (Δ*stk1*) in the presence of *stk2* (pDB31) or a control plasmid (pC194) were transformed with plasmid pAS10 expressing *pacK* under the control of Ptet promoter. Cells were grown to OD ≈ 0.2 and induced with aTc. After 2 hr of induction, cells were plated and colonies quantified (mean + SD, n = 3). Induction of *pacK* expression triggers cell death in the presence of *stk2* regardless of the presence or absence of *stk1*.

cells are killed by Stk2. It remains to be investigated whether this occurs by accelerating cell death, slowing down the phage, or some other mechanism. The phosphoproteome analysis performed here does not allow for differentiation of direct targets of Stk2 from targets phosphorylated by Stk1 as a result of Stk2 activation. Future work will focus on elucidating the molecular interaction between Stk2 and PacK, for which we only provide circumstantial evidence, as well as deciphering the exact phosphorylation cascade occurring during the response.

In sum, we provide strong evidence for a bacterial antiviral defense pathway involving a complex phosphorylation cascade and resulting in cell death through the modification of several essential cellular pathways. Other bacterial eSTKs have been shown to target different components of translation (EF-Tu and EF-P), transcription (various sigma and anti-sigma factors), cell division machinery (FtsZ), and central metabolism (Pereira et al., 2011), but none so far have been linked to phage defense. On the contrary, some phages have been described as using STKs in order to manipulate the host translation machinery for their own benefit (Robertson and Nicholson, 1992). It is also interesting to note that STKs play critical roles in the antiviral defense of eukaryotes. In particular, there are striking similarities between Stk2, the mammalian PKR, and the plant NIK1. All three STKs are activated by viral infection and target the translation machinery. Moreover, PKR not only inhibits the initiation of translation through phosphorylation of eIF-2α (Meurs et al., 1990), but can also trigger cell death through apoptosis (Dai et al., 2012; Hsu et al., 2004; Stark et al., 1998). Viral defense strategies that involve the serine/threonine phosphorylation of essential cellular pathways thus exist in both eukaryotes and bacteria. eSTKs have also recently been identified in archaea (Kennelly, 2014). In particular, the Ph0512p kinase from *Pyrococcus horikoshii OT3* was shown to phosphorylate the archaeal homolog of eIF2α (aIF2α) in vitro (Tahara et al., 2004). It is tempting to hypothesize that Ph0512p and other archaeal kinases could also be involved in viral defense, making this a universal strategy conserved across all domains of life.

EXPERIMENTAL PROCEDURES

Bacterial Strains and Growth Conditions

S. epidermidis LM1680 (Hatoum-Aslan et al., 2014), *S. aureus* RN4220 (Nair et al., 2011), and derivative strains were grown in tryptic soy broth (TSB) media at 37°C with shaking at 200 rpm. *S. epidermidis* LM1680 and *S. aureus* RN4220 were used as hosts for recombinant plasmids. Strains NCTC8325-4 and ST1004 (NCTC8325-4 Δstk1) are gifts from Michel Debarbouille. Chloramphenicol (10 μg/mL), erythromycin (10 μg/mL), and ampicillin (100 μg/mL) were added to the medium to prevent loss of plasmids derived from pC194, pE194 (Horinouchi and Weisblum, 1982a, 1982b), and pET15b (Novagen), respectively. *E. coli* BL21(DE3)AD494 (Novagen) was used for expression of recombinant proteins and grown in Luria-Bertani (LB) medium supplemented with 100 μg/mL ampicillin at 37°C.

Isolation of Phage CNPx

Phage CNPx (GenBank: KU-598975) was isolated as a plaque on a soft agar lawn of *S. epidermidis* LM1680 that was infected with phage CNPH82 (Daniel et al., 2007). LM1680 is resistant to phage CNPH82, and the isolation of the CNPx was a single fortuitous event that might have occurred via contamination with an environmental phage and recombination with CNPH82. Indeed, CNPx shares close to 100% homology with CNPH82 over 65% of its genome length, but carries a divergent segment of ~13 kb in the region of the genome containing the lysogenic operon and the early lytic genes.

Introduction of Plasmids in Staphylococci

Plasmid constructions are detailed in the Supplemental Experimental Procedures. Lists of plasmids and oligonucleotides used in this study are provided in Tables S5 and S6. After DNA assembly, all plasmids were first electroporated in *S. aureus* strain RN4220. Briefly, cells were grown to an optical density (600 nm) of 0.8 and washed three times in cold water and concentrated 100× in 10% glycerol. Electroporation of dialyzed DNA was performed in 2 mm cuvettes using the following settings: 100 Ω, 2.5 kV, 25 μF. In order to introduce plasmids in other Staphylococci strains, plasmids were extracted from RN4220 using the NucleoSpin Plasmid kit (Macherey Nagel) with the following modification: 4 μg lysostaphin (Ambi) was added to the A1 buffer, and cells were incubated 1 hr at 37°C in this buffer before resuming the protocol as described. Plasmids extracted from RN4220 can then be introduced in other Staphylococci through electroporation following the same protocol.

Overproduction and Purification of Stk2 and Derivatives

Plasmids pET15Ω*stk2*(*sa0077*) and derivative mutants were introduced into *E. coli* BL21λDE3/AD494. The transformants were grown in 1 L LB medium with shaking at 25°C until optical density $(OD)_{600}$ = 0.5, IPTG (0.5 mM) was added to induce protein production, and incubation was pursued for 6 hr at 25°C. Cells were then harvested by centrifugation at 3,000 g for 10 min. Since Stk2 and its mutants were not soluble and retained in inclusion bodies, a step of denaturation/renaturation using guanidine chloride according to London (London et al., 1974) and Goldberg (Goldberg et al., 1996) was performed before the purification on an Ni-NTA column.

In Vitro Phosphorylation Assay

Phosphotransfer to purified Stk2 and its derivatives was performed in a buffer containing 25 mM Tris-HCl (pH 7.5), 1 mM DTT, 2.5 mM MnCl2, 10 mM ATP,

and 5 μCi [γ-32P]-ATP and incubated at 37°C 10–30 min following the substrate. The reaction was stopped by the addition of 20% Laemmli 5X (Sigma), followed by electrophoresis on SDS-PAGE and autoradiography.

Phosphoamino Acid Analysis

The method used to detect acid-stable phosphoamino acids was described previously (Duclos et al., 1991).

Phage Production

Phages were mixed with *S. aureus* RN4220 in soft tryptic soy agar (TSA) supplemented with CaCl2 (5 mM) and then poured on top of TSA plate supplemented with CaCl2 (5 mM). The plates were incubated overnight at 37°C. Soft TSA lawns were then resuspended in PBS solution (1×) and centrifuged, and the lysate containing the phage was filtered on a 0.22 μM filter. To measure phage titers, serial dilutions were spotted on a soft agar lawn of RN4220 in TSA supplemented with CaCl2 (5 mM), and plaque-forming units (PFUs) were quantified after incubation overnight at 37°C.

EOP Assays

Phage lysates containing ~10^7 PFU/μL CNPX, phage 80alpha, phage 85, phiNM1, phiNM2, phiNM4, or Staph1N were serially diluted and spotted on soft TSA lawns supplemented with 5 mM CaCl2 and containing either *S. epidermidis* LM1680 or *S. aureus* RN4220, TB4, 8325-4, or 8325-4(Δ*stk1*) cells containing the indicated plasmids. PFUs were quantified after incubation overnight at 37°C.

ECOI

RN4220 cells carrying plasmid pC194 or pDB31 were grown to an OD of 0.6 and incubated 10 min at 37°C with phage phiNM1 at an MOI of 5. Cells were then washed twice in fresh TSB to remove unbound phages and plated on a lawn of RN4220 cells. ECOI was obtained by dividing the number of plaques (or center of infections) obtained after infecting cells carrying pDB31 by the number of plaques obtained with cells carrying pC194.

Phage DNA Isolation and Sequencing

Samples of phage lysates were treated with DNase and RNase to a final volume of 200 μL for 30 min at 37°C followed by treatment with EDTA (pH 8.0, 5 mM) and Proteinase K (0.5 mg/mL) for 30 min at 37°C. Phage DNA was then purified using a PCR purification kit (Macherey Nagel). Phage DNA was sequenced using the Nextera library preparation kit from Illumina and sequenced on an MiSeq device.

Adsorption Assay

Recipient RN4220 cells were grown to an OD of 0.6 and incubated with phage phiNM1 at an MOI of 1 for 10 min. Cells were then centrifuged and the number of phages remaining in the lysates was quantified (n_{ad}). Adsorption efficiency is computed as ($1 - n_{ad}/n_{tot}$), were n_{tot} is the total number of phages added to the sample, and reported as percentages.

Growth Curves

S. aureus strains (RN4220, RN4220/pC194, RN4220/pDB31, RN4220/pDB275, and RN4220/pFD6) were grown in triplicate overnight at 37°C and diluted 1:100 in 200 μL TSB broth in a 96-well microplate that was incubated at 37°C with shaking in an Infinite M200 PRO reader (TECAN). Absorbance was measured at 600 nm every 10 min. For RN4220, RN4220/pC194, and RN4220/pDB31, after 1 hr of growth (OD_{600} ≈ 0.2), 10 μL phiNM1 (4.10^7 PFU/μL) phage was added. For RN4220, RN4220/pDB275, and RN4220/pFD6, when OD_{600} reached 0.2, the cultures were induced by aTc for 1 hr and then phage 80alpha (5.10^7 PFU/μL) was added.

Prophage Induction

S. aureus strains (RN4220, RN4220::phiNM1, RN4220::phiNM1/pC194, and RN4220::phiNM1/pDB31) were grown in triplicate overnight at 37°C, diluted 1:100 in TSB, and incubated at 37°C with shaking. When cultures reached OD_{600} ≈ 0.4, mitomycin C was added at a final concentration of 2 μg/mL. After 3 hr of incubation in the presence or in the absence of mitomycin C, the samples were serially diluted and plated to quantify the number of surviving bacteria. Samples were also centrifuged to recover the supernatant and measure the phage titer.

Induction of Candidate Activators of Stk2

S. aureus strains were grown in triplicate overnight at 37°C, diluted 1:100 in TSB, and incubated at 37°C with shaking. When cultures reached OD_{600} ≈ 0.2, aTc was added at a final concentration of 0.5 μg/mL. All the strains were grown in parallel without aTc as a control. After 1.5 hr of incubation in the presence or absence of aTc, the samples were serially diluted and 5 μL was spotted in TSA with appropriate antibiotics to count viable bacteria.

Mass Spectrometry

Mass spectrometry methods are detailed in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The GenBank accession number for the sequence of the phage CNPx reported in this paper is GenBank: KU598975.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.chom.2016.08.010.

AUTHOR CONTRIBUTIONS

F.D., J.-P.D., A.S., and D.B. performed the experiments. B.D., F.D., and D.B. wrote the manuscript. H.M. performed the mass spectrometry experiments and analyzed the data. A.B. performed the phylogenetic analysis.

ACKNOWLEDGMENTS

We are indebted to Dr. Luciano Marraffini for reagents and support, to the group of Dr. Romain Koszul for its help with phage sequencing, and to Michel Debarbouille for useful discussions and strain gifts. This study has received funding from the French government's Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant no. ANR-10-LABX-62-IBEID).

Received: November 16, 2015
Revised: July 6, 2016
Accepted: August 29, 2016
Published: September 22, 2016

REFERENCES

Abedon, S.T. (2012). Bacterial 'immunity' against bacteriophages. Bacteriophage *2*, 50–54.

Bae, T., Baba, T., Hiramatsu, K., and Schneewind, O. (2006). Prophages of Staphylococcus aureus Newman and their contribution to virulence. Mol. Microbiol. *62*, 1035–1047.

Beltramini, A.M., Mukhopadhyay, C.D., and Pancholi, V. (2009). Modulation of cell wall structure and antimicrobial susceptibility by a Staphylococcus aureus eukaryote-like serine/threonine kinase and phosphatase. Infect. Immun. *77*, 1406–1416.

Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A.J.R. (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat. Protoc. *4*, 484–494.

Chopin, M.C., Chopin, A., and Bidnenko, E. (2005). Phage abortive infection in lactococci: variations on a theme. Curr. Opin. Microbiol. *8*, 473–479.

Dai, R., Yan, D., Li, J., Chen, S., Liu, Y., Chen, R., Duan, C., Wei, M., Li, H., and He, T. (2012). Activation of PKR/eIF2α signaling cascade is associated with dihydrotestosterone-induced cell cycle arrest and apoptosis in human liver cells. J. Cell. Biochem. *113*, 1800–1808.

Daniel, A., Bonnen, P.E., and Fischetti, V.A. (2007). First complete genome sequence of two Staphylococcus epidermidis bacteriophages. J. Bacteriol. *189*, 2086–2100.

Débarbouillé, M., Dramsi, S., Dussurget, O., Nahori, M.A., Vaganay, E., Jouvion, G., Cozzone, A., Msadek, T., and Duclos, B. (2009). Characterization of a serine/threonine kinase involved in virulence of Staphylococcus aureus. J. Bacteriol. *191*, 4070–4081.

Didier, J.P., Cozzone, A.J., and Duclos, B. (2010). Phosphorylation of the virulence regulator SarA modulates its ability to bind DNA in Staphylococcus aureus. FEMS Microbiol. Lett. *306*, 30–36.

Donat, S., Streker, K., Schirmeister, T., Rakette, S., Stehle, T., Liebeke, M., Lalk, M., and Ohlsen, K. (2009). Transcriptome and functional analysis of the eukaryotic-type serine/threonine kinase PknB in Staphylococcus aureus. J. Bacteriol. *191*, 4056–4069.

Duclos, B., Marcandier, S., and Cozzone, A.J. (1991). Chemical properties and separation of phosphoamino acids by thin-layer chromatography and/or electrophoresis. Methods Enzymol. *201*, 10–21.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.

Friedman, D.I., Mozola, C.C., Beeri, K., Ko, C.C., and Reynolds, J.L. (2011). Activation of a prophage-encoded tyrosine kinase by a heterologous infecting phage results in a self-inflicted abortive infection. Mol. Microbiol. *82*, 567–577.

Goldberg, M.E., Expert-Bezançon, N., Vuillard, L., and Rabilloud, T. (1996). Non-detergent sulphobetaines: a new class of molecules that facilitate in vitro protein renaturation. Fold. Des. *1*, 21–27.

Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633–637.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., and Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. *34*, 169–183.

Hatoum-Aslan, A., Maniv, I., Samai, P., and Marraffini, L.A. (2014). Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. J. Bacteriol. *196*, 310–317.

Horinouchi, S., and Weisblum, B. (1982a). Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. J. Bacteriol. *150*, 815–825.

Horinouchi, S., and Weisblum, B. (1982b). Nucleotide sequence and functional map of pE194, a plasmid that specifies inducible resistance to macrolide, lincosamide, and streptogramin type B antibodies. J. Bacteriol. *150*, 804–814.

Hoskisson, P.A., Sumby, P., and Smith, M.C.M. (2015). The phage growth limitation system in Streptomyces coelicolor A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. Virology *477*, 100–109.

Hsu, L.C., Park, J.M., Zhang, K., Luo, J.L., Maeda, S., Kaufman, R.J., Eckmann, L., Guiney, D.G., and Karin, M. (2004). The protein kinase PKR is required for macrophage apoptosis after activation of Toll-like receptor 4. Nature *428*, 341–345.

Huse, M., and Kuriyan, J. (2002). The conformational plasticity of protein kinases. Cell *109*, 275–282.

Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R., and Marraffini, L.A. (2013). Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. PLoS Genet. *9*, e1003844.

Kanneganti, T.D. (2010). Central roles of NLRs and inflammasomes in viral infection. Nat. Rev. Immunol. *10*, 688–698.

Kennelly, P.J. (2014). Protein Ser/Thr/Tyr phosphorylation in the Archaea. J. Biol. Chem. *289*, 9480–9487.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat. Rev. Microbiol. *8*, 317–327.

Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P., and Jørgensen, T.J.D. (2005). Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol. Cell. Proteomics *4*, 873–886.

Łobocka, M., Hejnowicz, M.S., Dąbrowski, K., Gozdek, A., Kosakowski, J., Witkowska, M., Ulatowska, M.I., Weber-Dąbrowska, B., Kwiatek, M., Parasion, S., et al. (2012). Genomics of staphylococcal Twort-like phages—potential therapeutics of the post-antibiotic era. Adv. Virus Res. *83*, 143–216.

Lomas-Lopez, R., Paracuellos, P., Riberty, M., Cozzone, A.J., and Duclos, B. (2007). Several enzymes of the central metabolism are phosphorylated in Staphylococcus aureus. FEMS Microbiol. Lett. *272*, 35–42.

London, J., Skrzynia, C., and Goldberg, M.E. (1974). Renaturation of Escherichia coli tryptophanase after exposure to 8 M urea. Evidence for the existence of nucleation centers. Eur. J. Biochem. *47*, 409–415.

Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J. Bacteriol. *193*, 6039–6056.

Marchler-Bauer, A., and Bryant, S.H. (2004). CD-Search: protein domain annotations on the fly. Nucleic Acids Res. *32*, W327–W331.

Meurs, E., Chong, K., Galabru, J., Thomas, N.S.B., Kerr, I.M., Williams, B.R.G., and Hovanessian, A.G. (1990). Molecular cloning and characterization of the human double-stranded RNA-activated protein kinase induced by interferon. Cell *62*, 379–390.

Molle, V., Zanella-Cleon, I., Robin, J.P., Mallejac, S., Cozzone, A.J., and Becchi, M. (2006). Characterization of the phosphorylation sites of Mycobacterium tuberculosis serine/threonine protein kinases, PknA, PknD, PknE, and PknH by mass spectrometry. Proteomics *6*, 3754–3766.

Nair, D., Memmi, G., Hernandez, D., Bard, J., Beaume, M., Gill, S., Francois, P., and Cheung, A.L. (2011). Whole-genome sequencing of Staphylococcus aureus strain RN4220, a key laboratory strain used in virulence research, identifies mutations that affect not only virulence factors but also the fitness of the strain. J. Bacteriol. *193*, 2332–2335.

Novick, R.P. (1989). Staphylococcal plasmids and their replication. Annu. Rev. Microbiol. *43*, 537–565.

Palm, N.W., and Medzhitov, R. (2009). Pattern recognition receptors and control of adaptive immunity. Immunol. Rev. *227*, 221–233.

Pereira, S.F.F., Goss, L., and Dworkin, J. (2011). Eukaryote-like serine/threonine kinases and phosphatases in bacteria. Microbiol. Mol. Biol. Rev. *75*, 192–212.

Pumplin, N., and Voinnet, O. (2013). RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. Nat. Rev. Microbiol. *11*, 745–760.

Robertson, E.S., and Nicholson, A.W. (1992). Phosphorylation of Escherichia coli translation initiation factors by the bacteriophage T7 protein kinase. Biochemistry *31*, 4822–4827.

Sadler, A.J., and Williams, B.R.G. (2008). Interferon-inducible antiviral effectors. Nat. Rev. Immunol. *8*, 559–568.

Stark, G.R., Kerr, I.M., Williams, B.R.G., Silverman, R.H., and Schreiber, R.D. (1998). How cells respond to interferons. Annu. Rev. Biochem. *67*, 227–264.

Sumby, P., and Smith, M.C.M. (2002). Genetics of the phage growth limitation (Pgl) system of Streptomyces coelicolor A3(2). Mol. Microbiol. *44*, 489–500.

Tahara, M., Ohsawa, A., Saito, S., and Kimura, M. (2004). In vitro phosphorylation of initiation factor 2 alpha (aIF2 alpha) from hyperthermophilic archaeon Pyrococcus horikoshii OT3. J. Biochem. *135*, 479–485.

Truong-Bolduc, Q.C., Ding, Y., and Hooper, D.C. (2008). Posttranslational modification influences the effects of MgrA on norA expression in Staphylococcus aureus. J. Bacteriol. *190*, 7375–7381.

Yan, N., and Chen, Z.J.J. (2012). Intrinsic antiviral immunity. Nat. Immunol. *13*, 214–222.

Young, T.A., Delagoutte, B., Endrizzi, J.A., Falick, A.M., and Alber, T. (2003). Structure of Mycobacterium tuberculosis PknB supports a universal activation mechanism for Ser/Thr protein kinases. Nat. Struct. Biol. *10*, 168–174.

Zorzatto, C., Machado, J.P.B., Lopes, K.V.G., Nascimento, K.J.T., Pereira, W.A., Brustolini, O.J.B., Reis, P.A.B., Calil, I.P., Deguchi, M., Sachetto-Martins, G., et al. (2015). NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. Nature *520*, 679–682.