TÉCNICO LISBOA

Cas4 solo in phages enhances host CRISPR autoimmunity

Cátia Sofia Marques Pereira

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisor(s)

Prof. Stan Brouns

Prof. Miguel Nobre Parreira Cacho Teixeira

Examination Committee

Chairperson: Prof. Jorge Humberto Gomes Leitão

Supervisor: Prof. Miguel Nobre Parreira Cacho Teixeira

Member of the Committee: Dr. Cláudia Sofia Pires Godinho

October 2018

The work presented in this thesis was performed at the Department of Bionanoscience of Technische Universiteit Delft (Delft, Netherlands), during the period March-September 2018, under the supervision of Prof. Stan Brouns and Dr. Cristóbal Almendros, and within the frame of the Erasmus programme. The thesis was co-supervised at Instituto Superior Técnico by Prof. Miguel Teixeira.

Abstract

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) together with *cas* (CRISPR associated) genes constitute an adaptive immune system found in prokaryotes. Microbes evolved a vast diversification of these systems that can be classified in two major classes and more than 30 subtypes according to their *cas* genes content. However, despite of their large diversity, in all CRISPR-Cas systems the immunological memory is adapted by integrating small fragments of foreign DNA (protospacers) into the CRISPR locus. Subsequently, the CRISPR array is transcribed resulting in short CRISPR RNAs (crRNAs) that will later guide the Cas proteins (encoded by the *cas* genes) to cleave and destroy the invading genetic elements.

One of the CRISPR associated proteins is Cas4 which role was recently described. The *cas4* gene is usually located next to the *cas1* or *cas2* in different systems. The association of Cas4 with the Cas1 and Cas2, constitute the CRISPR acquisition machinery that is crucial in the recognition, processing and orientation of protospacer integration. Curiously, *cas4* genes can also be found not associated with the CRISPR-cas loci in some bacterial and *archaeal* genomes as well as plasmids or bacteriophages, being their role unknown.

Here, was studied the phylogenomics of Cas4 solo in phages (vCas4) and their influence in CRISPR adaptation through *in vivo* and *in vitro* assays. Was demonstrated that, notwithstanding the vCas4 does not interact with the CRISPR acquisition module, the rates of novel spacers acquired decrease. Moreover, the sequencing of those new spacers revealed an enrichment of host genome derived spacers, which would contribute to CRISPR autoimmunity.

Keywords: CRISPR-Cas system, Cas proteins, sCas4, vCas4, bacteriophages, phylogenomics, CRISPR adaptation, type I-E CRISPR-Cas systems, type I-C CRISPR-Cas systems, protein activity

Resumo

A combinação de CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) com os seus genes associados (genes *cas*), constitui um dos sistemas imunitários adaptativos que pode ser encontrado em procariontes. Estes organismos desenvolveram uma vasta diversidade destes sistemas, que podem ser divididos em duas classes e mais de 30 subtipos, consoante o seu conteúdo em *cas* genes. A capacidade de adaptação da sua memória imunológica é conseguida através da integração de pequenos fragmentos de DNA dos organismos invasores (*protospacers*) no CRISPR *locus*. O CRISPR *array* é, depois, transcrito levando à formação de pequenos CRISPR RNAs (crRNAs) que recrutam as proteínas Cas (codificadas pelos genes *cas*) e levam, por conseguinte, à destruição do material genético dos organismos invasores.

Uma das proteínas CRISPR associadas é Cas4, cuja função foi recentemente descrita. O gene *cas4* encontra-se localizado, em diversos sistemas CRISPR, nas proximidades dos genes *cas1* ou *cas2*. A associação de Cas4 com Cas1 e Cas2 constitui a maquinaria necessária à acquisição em sistemas CRISPR, sendo crucial no reconhecimento, processamento e orientação de protospacers, durante a sua integração. Curiosamente, os genes *cas4* foram recentemente encontrados não associados ao CRISPR-Cas loci, em genomas de bactérias e *archaea*, bem como plasmídeos e bacteríofagos, sendo o seu papel desconhecido.

Aqui, estudou-se a filogenia desses Cas4 solo em fagos (vCas4) e a sua influência na adaptação dos sistemas cRISPR através de ensaios *in vivo* and *in vitro* tendo sido demonstrado que esta proteína, embora não interaja diretamente com o módulo de acquisição dos sistemas CRISPR, influencia os rácios de acquisição de novos spacers. Além disso, por sequenciaçao dos novos spacers adquiridos, foi possível revelar que mais spacers com origem no organismo hospedeiro são integrados o que, consequentemente, contribui para a auto-imunidade dos sistemas CRISPR.

Keywords: sistema CRISPR-Cas, proteínas Cas, sCas4, vCas4, bacteriófagos, filogenia, adaptação CRISPR, sistemas CRISPR-Cas tipo I-E, sistemas CRISPR-Cas tipo I-C, actividade proteica

Acknowledgements

First, I would like to thank Dr. Stan Brouns, for putting your trust in me and giving me the opportunity to work in your lab. Then, to the person that made it possible: Cristóbal. Which advice and knowledge were fundamental to build up this project. For teaching me everything I needed to complete this thesis and for all the great input. Thank you! Not only for the technical advice but also for the help in building me as a future scientist. And for the most important thing: never letting me be less than the best I could and always asking me more.

To all the Stan Brouns lab members: Seb, Becca, Rita, Franklin, Tóbal, Jochem, Cristian, Teunke and Anna and, of course, the student's team: Ana, Anwar, Jasper, Jeroen, Rik, Amanda and Hielke.

To Seb, for all the primers stock. My PCRs wouldn't be possible without you (literally). For being ready to help in case of a missing supervisor or, actually, in any case.

To Franklin, Rita and Ana: my portuguese family in Brouns lab. To Rita for being as a mom in Delft and for the big heart. To Frankin for, besides all the bioinformatic help and nice input, fragilizing my wall and for teaching me that not controlling everything is also fine. To Ana, for the support inside and outside the lab. For being the best partner and for making me feel like home since the first day (or maybe just the second).

To Eric, the best lab-buddy ever. For the music, the support, for being always there and for the best breaks outside. Having you in the lab made everything easier. To Anwar, for the good talks and for making part of the never failing 'Bourgeoise' team.

To my Portuguese fellows: Adriana, Bruno and Kets. For the afternoon pauses even when in a bad mood. To Bruno and Kets for suggesting that maybe I was a bit aggressive. Life changing. For the evenings writing and for the unconditional support during this experience. A very special note to Adriana without whom this months would not be the same. For being the best flatmate, roommate and 'bedmate'. I couldn't ask for better partner to live this adventure. Thank you for these amazing six months but also for the last five years and the certainty that this is never-ending friendship.

To Haris, a note of heartfelt gratitude. For kindly reading and reviewing this thesis and for the genuine interest on it. For teaching me that life and human relations can be so easy, the pleasures of relaxing and for one of the prettiest occasional things in my life: meeting you. For the three weeks we were together in Delft and for, somehow, being always present, supporting me and making this thesis possible.

To the Erasmus team: Adriana, Bruno, Daniel, Diogo, Félix, George, May, Judith, Kets, Pietro, Diogo, Thea and all the others that crossed my path. For making my experience in Delft even better. Without forgetting the native dutch: Rogier. For the incredibly good introduction to the dutch culture.

Couldn't finish with mentioning João. For the amazing years we spent together, fundamental to bring me here. It would not be possible without you. For the support, the experiences and all the barriers we crossed together. It was amazing to grow up with you.

A special word also for Tiago. For being the best friend ever and a big part of my life in the last five years and for the cooperation in every single project or report. Also for the awesome visit in Delft.

To all my remaining friends from Portugal. In special to Mariana and Alexandra, for also visiting me.

Finally, to my parents without whom this would not be (literally) possible. Obrigada por me fazerem sonhar sempre mais alto e alimentarem sempre cada um desses sonhos. Por estarem sempre presentes. É incrível ir mas ainda melhor saber que, onde quer que vá, vocês estarão por perto. Pelo amor incondicional e pelo apoio incrível em todos os momentos.

À minha irmã, que me eleva a fasquia desde que nasci. Obrigada por teres dos percursos profissionais mais incríveis que conheço e nunca teres deixado qualquer espaço para querer ser menos que a melhor versão de mim própria.

Aos três: pai, mãe e irmã por serem co-autores deste projeto que é viver.

Lastly, to Atum, the cat. For still loving me after six months of being an orphan and for being the best and most loyal partner of all my adventures.

Contents

List of Acronyms and Abreviations

AddB ATP-dependent helicase/deoxyribonuclease subunit B

BLAST Basic Local Alignment Search Tool

Cas CRISPR associated proteins

CRISPR Clustered Regularly Interspaced Palindromic Repeats

CRT CRISPR Recognition Tool

crRNA CRISPR RNAs

DNA Deoxyribonucleic Acid

dsDNA Double-stranded DNA

DTT Dithiothreitol

Fw Forward primer

HEPES 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HF High Fidelity

His6 Polyhistidine

IDT Integrated DNA Technologies

IPTG Isopropyl beta-D-thiogalactopyranoside

LIC Ligation-Independent Cloning

LB Luria-Bertani media

LBA Luria-Bertani media supplemented with agar

MES 2-(N-morpholino)ethanesulfonic acid

MGE Mobile Genetic Elements

MS Mass Spectrometry

Ni-NTA Nickel-Nitrilotriacetic acid

PAGE Polyacrylamide Gel Electrophoresis

PAM Protospacer Adjacent Motif

RAxML Randomized Axelerated Maximum Likelihood

RNA Ribonucleic Acid

Rv Reverse primer

SF Small Fragments

ssDNA Single-stranded DNA

SUMO Small Ubiquitin-like Modifier

Ta Annealing temperature

TAE Tris-acetate-EDTA

TBE Tris-borate-EDTA

Tm Melting temperature

Tris tris(hydroxymethyl)aminomethane

TTV1 Thermoproteus tenax virus 1

PAM Protospacer Adjacent Motif

PCR Polymerase Chain Reaction

PNK T4 Polynucleotide Kinase

SEC Size-Exclusion Chromatography

T7 RNAP T7 RNA polymerase

X-Gal 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

Chapter 1

Introduction

Dynamic interactions between phages and hosts have been studied since the early days of molecular biology. Bacteria are outnumbered by a factor of 10 by phages that infect them - bacteriophages [27, 53, 57]. This phage-host arms race shaped the evolution of microbes to evolve a number of pathways to enable their infection by invader elements that compromise the fitness of the population [35]. The pressure caused by this ever-changing mobile genetic elements (MGEs) that include, not only bacteriophages, but also other entities such as conjugative plasmids, allowed for prokaryotes to have complex systems to thrive in hostile and competitive environments where both predator and prey fight and evolve for survival.

These interactions lead to the development of a vast diversification of defense mechanisms in prokaryotes that can be classified as innate or adaptive immune systems [3, 61]. The innate systems act as the first line of defense and include receptor modification, restriction modification systems and abortive infection systems [61]. If these systems fail and the viral replication outpaces innate defenses, the adaptive immune systems are activated as a second line of defense [3] showing a higher specificity. CRISPR and their associated *cas* genes encode one such adaptive immune system mechanism allowing cells to restrict incoming nucleic acids [20].

1.1 CRISPR-Cas

CRISPR-Cas system is an adaptive immune system that uses clustered regularly interspaced short palindromic repeats (CRISPR) and Cas (CRISPR-associated) proteins. Its earliest description dates back to the late 1980's with the discovery of a repeat array in *Escherichia coli* genomes [19] that were found also in other prokaryotes [21, 37]. These repeats were later denominated CRISPR and found to be associated with proteins that were thereafter termed CRISPR associated genes (*cas*) [21, 37]. The repeats are interspaced by sequences termed spacers that were found to have homology to foreign genetic elements such as bacteriophage genomes and conjugative plasmids [39]. Later studies revealed that the spacers were acquired by bacteria from these foreign genetic elements and incorporated in the CRISPR array [2] and that the content of this array has an influence in phage sensitivity [2].

Found in approximately 45% of bacteria and 85% of *archaea* [24], the CRISPR array consist of a cluster of a highly variable number of repeats with a particular sequence and length, interspaced by spacers and an AT-reach *leader* sequence [21, 37]. This sequence has two main roles: contains the promoter of the CRISPR array (fundamental in the production of pre-crRNA) and is recognized by the Cas1-Cas2 complex

for acquisition (see section CRISPR-Cas mechanism) . In addition to the CRISPR array, an operon of *cas* genes is usually found in close proximity. The Cas proteins encoded by these genes are responsible to provide the enzymatic machinery required by the system to work [21].

1.2 Diversity of CRISPR-Cas systems

The CRISPR-Cas systems have evolved a vast diversification being categorized by both their phylogenomics and *cas* genes content into two classes, six types and more than twenty subtypes [24]. Class 1 is divided into type I, III and IV and Class 2 in types II, V and VI (Figures 1 and 2).

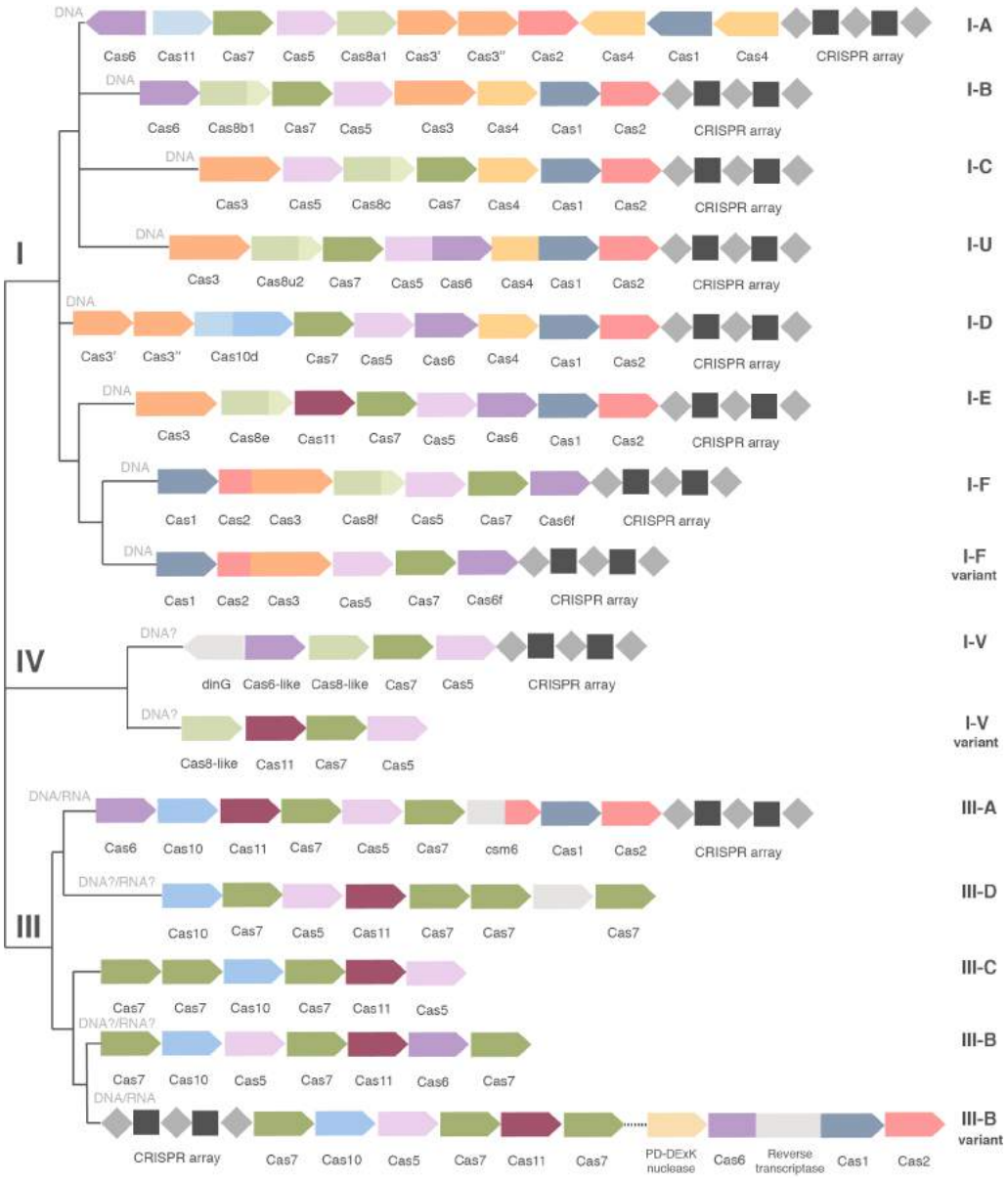

Figure 1: Classification of Class 1 CRISPR systems. Representation of Class 1 division in types I, III and IV and corresponding subtypes. For each subtype, the organization of the CRISPR-cas locus and the (predicted) target (DNA, RNA or both) is shown. Modified from Koonin *et al.*, 2017.

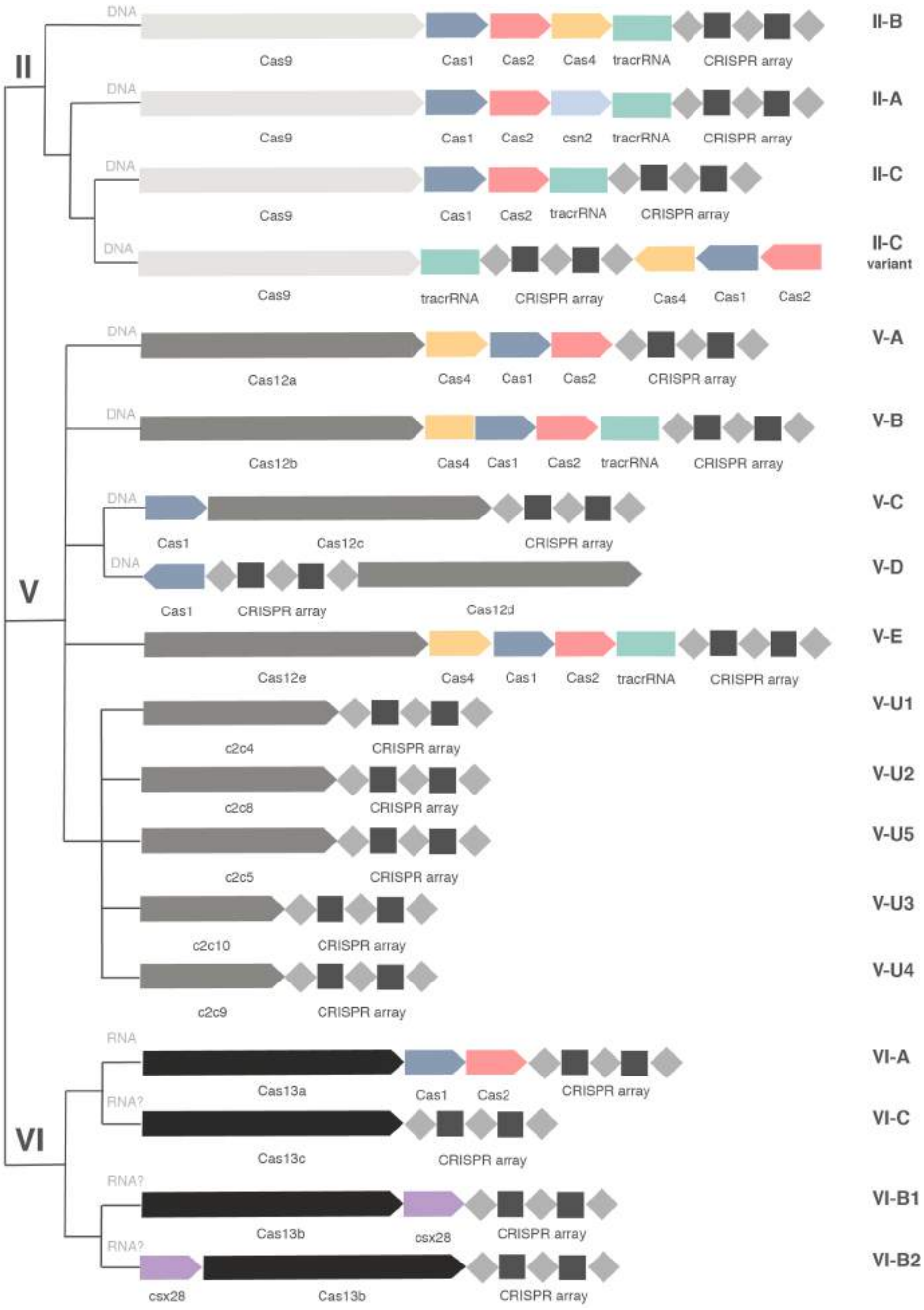

Figure 2: Classification of Class 2 CRISPR systems. Representation of Class 2 division in types II, V and VI and corresponding subtypes. For each subtype, The organization of the CRISPR-cas locus and the (predicted) target (DNA, RNA or both) is shown. Modified from Koonin *et al.*, 2017.

The integration of new spacers in the CRISPR array, encoded by *cas1* and *cas2*, is well conserved in all the six types of CRISPR systems [24]. They differ in the effector complex that mediates the destruction of foreign DNA which is functionally distinct depending on the CRISPR type [35]. Except for type IV systems, it was experimentally characterized that types I, II and V systems target DNA whereas type VI systems target RNA and type III systems target both DNA and RNA [24].

The systems investigated in this study were from class 1 type I which is the most abundant type in nature, found in around 50% of all bacteria and *archaea* in which a CRISPR loci was identified [33]. The subtypes studied were the I-E and I-C.

1.3 CRISPR-Cas Mechanism

Despite the high diversity of CRISPR-Cas systems, their mechanism can be encompassed into three main stages: adaptation, expression and processing and, finally, interference (Figure 3).

1.3.1 Adaptation

Adaptation is the first stage of the CRISPR-Cas mechanism in which new spacers are acquired by the system in order to update the repertoire of recognized foreign invaders (Figure 3). This way, a small fragment of DNA, termed protospacer, is acquired from the MGE [2] and integrated in the CRISPR array, forming a new spacer. These spacer fragments have variable size, depending on the CRISPR system [2].

To avoid acquisition of spacers from the host DNA that would lead to autoimmunity, the CRISPR systems use the DNA machinery repair of the hosts to generate protospacers that uses the RecBCD or AddB machinery in the cases of Gram-negative or Gram-positive organisms, respectively [35]. The RecBCD molecules bind to the free ends of dsDNA and performs end resection during homologous recombination that consequently stimulates the acquisition of spacers from double-strand breaks [40]. However, this activity is slowed by the presence of an eight nucleotide sequence motif (*chi* motifs) that are enriched in the host chromosome when compared with phages or plasmids [32]. This way, since the host chromosome has more *chi* motifs, less spacers are acquired from there.

Cas1 is the most conserved Cas protein, present in all of the six types of CRISPR systems [24]. In the context of CRISPR immunity, this protein interacts with Cas2 forming a complex responsible for spacer integration (hereafter, Cas1-Cas2) [43, 70]. This complex has two separate DNA-binding proteins mediating the connection between the incoming protospacer and the CRISPR array [67]. The incorporation of spacers is possible due to the two nucleophilic attack reactions that are catalyzed by this complex once loaded with the incoming spacer [42]. The first reaction allows the ligation between the leader end and the first repeat of the CRISPR array and the second allows the ligation of the spacer end to the repeat sequence [42]. These reactions are responsible not only for the incorporation of new spacers but also for repeat duplication. This is possible since the terminal 3'-OH of each strand of the protospacer carries out a nucleophilic attack reaction on each end of the repeat sequence [42,68]. These reactions produce an intermediate in which the 3' ends of a dsDNA protospacer are ligated to a ssDNA repeat sequence that is further completed and ligated [66, 68]. From this process results a CRISPR array with a newly integrated spacer and duplicated repeat sequences directly after the leader (the promoter of the CRIPSR array that allows its transcription).

The new spacers acquired by the Cas1-Cas2 complex are predominately incorporated in the leader end of the CRISPR array [70]. This way, by its organization, it is possible to have a record of past infections since newer memories are located at the leader end and the most ancestral spacers are positioned in the trailer end [70]. This chronological organization of spacers optimizes the immune response of the CRISPR system since the spacers positioned near the leader end provide more robust immunity responses when compared to the more downstream positioned ones. This is caused by phenomena such as the differential expression of crRNAs across the CRISPR array [35]. How polarized addition of new spacers is achieved differs by CRISPR type and factors encoded by the host genome such as the integration host factors present in type I-E CRISPR systems that can be required for site-specific integration [35].

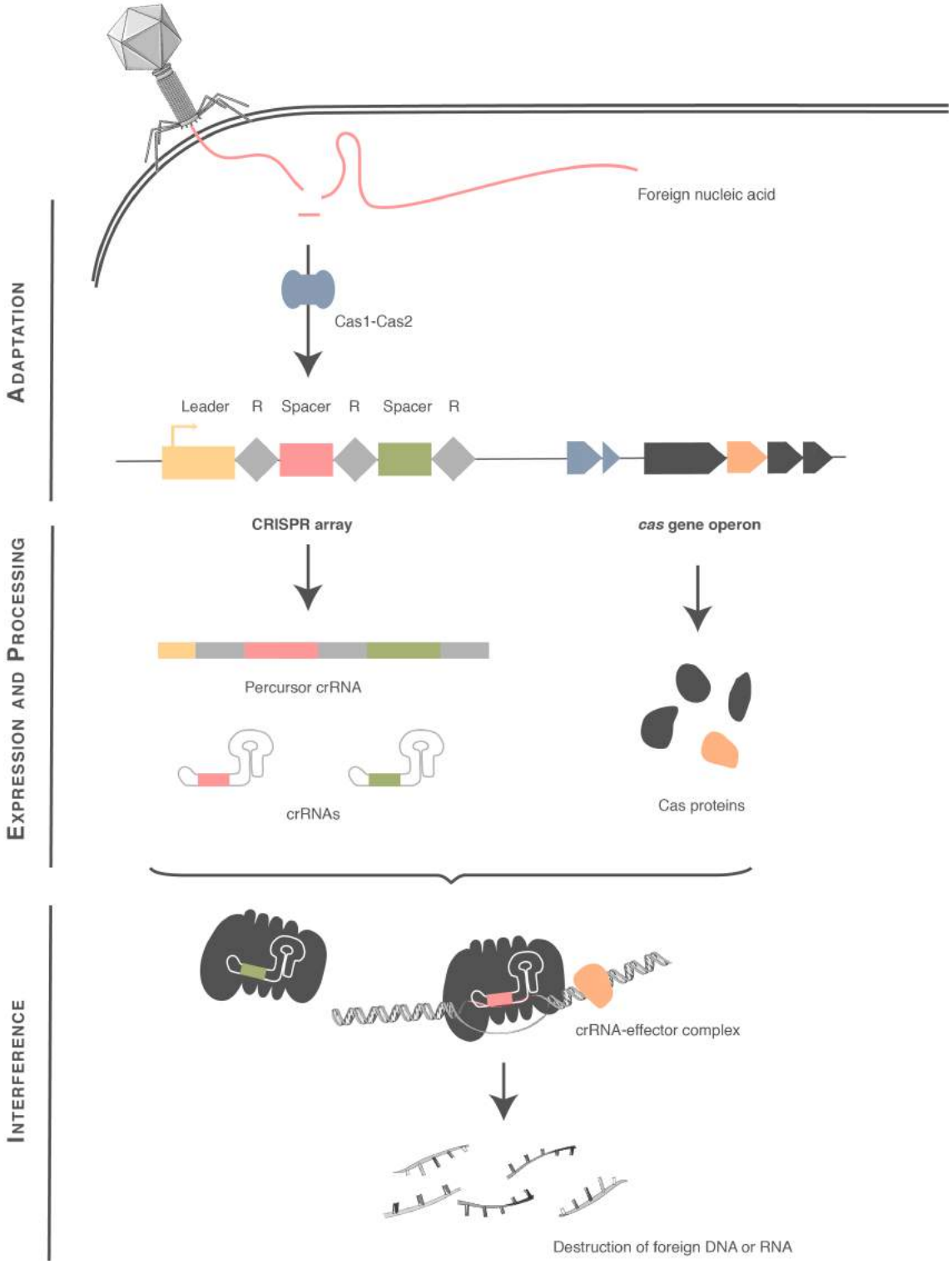

Figure 3: Schematic representation of CRISPR-Cas immune system mechanism During adaptation, the Cas1-Cas2 complex select and incorporate new spacers in the CRISPR array. During expression and processing, the CRISPR array is transcribed to produce crRNA that forms the crRNA-effector complex by binding Cas proteins. During interference, the foreign DNA is recognized and degraded. Modified from Jackson *et al.*, 2017.

The identification of convenient protospacers is based on the presence of a protospacer-adjacent motif (PAM) [38, 63]. This motif preceding the protospacer sequence ensures its correct orientation in the CRISPR array [59]. Thereby, the PAM plays an important role in both adaptation and interference stages of the

CRISPR mechanism since it allows for proper recognition of the target during interference [20, 62] (see Interference).

Other Cas protein that is known to be implicated in the adaptation phase in several subtypes of CRISPR-Cas systems is the Cas4 protein. It has so far been identified in the subtypes I-A, I-B, I-C, I-D and I-U of Class 1 CRISPR systems (see Figure 1) and subtypes II-B, V-A, V-B and V-E of Class 2 (see Figure 2) [24, 33].

In the CRISPR loci the *cas4* genes were found either adjacent to *cas1* and *cas2* or, in some cases, fused with *cas1* [24]. This fusion of *cas4* with *cas1* was the first suggestion that Cas4 proteins would be involved in the adaptation stage of the CRISPR-Cas mechanism [18]. Later, supporting that hypothesis, it was found that in subtype I-A loci, *cas1*, *cas2* and *cas4* genes form a single operon [46]. Recently, it was experimentally demonstrated that Cas4 crucial in the recognition, processing and orientation of protospacers integration [23, 28, 51, 55, 72].

Cas4 protein is a DNA nuclease containing a Fe-S cluster-binding module [30, 72]. It shows homology to nuclease motifs of proteins known to be involved in both recombination and repair processes in bacteria and eukaryotes such as RecB, AddB and Dna2 [18]. Moreover, it is demonstrated to be involved in the processing of 3'ssDNA overhangs in the protospacers, facilitating their incorporation in the CRISPR array [18, 28, 51]. Regarding structural characterization, early biochemical studies have described different Cas4 proteins as monomers, dimers and decamers [30, 72].

1.3.2 Expression and Processing

The adaptation stage of CRISPR-Cas mechanism is followed by the expression phase in which the CRISPR array and CRISPR locus are transcribed [6] (Figure 3).

The transcription of the CRISPR array leads to the formation of a long precursor CRISPR-RNA molecule (pre-crRNA) containing the repeat and the interspacing spacers [6]. This molecule is further processed forming hairpin-like structures due to the palindromic inverted repeat sequences present in the repeats [15]. This structure is recognized by a metal-independent endoribonuclease that is responsible for cleaving the pre-crRNAs within each repeat to produce mature crRNAs [8]. These molecules have a well defined structure and, in type I-E CRISPR-Cas systems is known to be formed by a 5' handle with 8 nucleotides, a 32 bp spacer sequence and a 21 nucleotides hairpin shaped 3'handle [34]. The homolog endoribonuclease is also responsible for capturing the mature crRNAs and assemble the effector complex Cascade (CRISPR associated complex for anti-viral defense) [22, 65].

1.3.3 Interference

Interference is the last stage of the CRISPR-Cas immune system mechanism (Figure 3). This multi-step process starts with the initial recognition of the invading sequences and, after target binding, finalizes with obstruction of nucleic acid invasion by target destruction [34].

Once generated, crRNAs use their base-pairing potential and serve as guides for the recognition of invasive targets [34]. This recognition process is performed by the effector complex that probes all the DNA for the correct three nucleotide PAM sequence [38]. Even if specific for the the right PAM, the effector complex

allows some variation in this sequence and multiple PAMs are capable of inducing direct interference [29,69].

After PAM scanning and recognition, the crRNA binds to the base-pairing target in a process termed R-loop formation [52]. It starts with the binding of the seed region (the first nucleotides of spacer sequence into the crRNA) to the target followed by the matching of the remaining crRNA molecule in a rolling capacity [52]. Mismatches in any step cause inhibition of R-loop formation and consequent termination of the process (non interference) [60].

In the cases where the crRNA fully matches the target nucleic acid, the Cas proteins undergo structural changes, that increases stabilization of the binding between the effector complex and the invader sequence [5]. Consequently, nuclease Cas proteins are recruited to the side promoting the final destruction of the invaders [14,26,64].

1.3.4 Cas4 Solo

The *cas4* gene is usually located next to the *cas1* or *cas2* in different systems however it can also be found not associated with the CRISPR-cas loci in some prokaryotic genomes as well as plasmids and bacteriophages. Recently, a phylogenomic study of Cas4 family nucleases was performed by Hudaiberdiev *et al.*, 2017. In this study, were analyzed the sequence profiles of Cas4 homologs being found a total of 7060 Cas4 homolog protein sequence, 883 from complete bacterial and archaeal genomes and 272 from viruses. From these sequences, a sequence similarity dendogram was constructed and, as expected, the Cas4 clustered in two main groups comprising the Cas4 proteins belonging to the COG1468 and COG4343 families that incorporate the majority of Cas4 known to be associated with CRISPR systems and the remaining Cas4 homologs were classified in three major groups based on their genomic context: CRISPR-Cas associated Cas4, Cas4 associated with MGEs and viruses, and solo-Cas4. With the phylogenomic study previously described it was then possible detect Cas4 proteins encoded in phage genomes forming specific and isolated clusters that were mostly similar to Cas4 proteins associated with type I CRISPR-Cas systems [18]. These vCas4 were under the scope of some other recent studies [16,25] however, this phylogenomic study performed by Hudaiberdiev *et al.*, 2017 have shown that these proteins have high diversity being present in a wide rage of phage genomes. This suggests that the solo-Cas4 protein encoded in the phage genome could be a CRISPR-associated Cas4 protein captured by the ancestral virus [18]. Moreover, were identified Cas4 homologs of numerous phages from a large, well-suported clade which includes *Cyanobacteria* and *Proteobacteria*, suggesting the dissemination of the *cas4* genes among phages [18].

1.4 Aims of this study

Until now Cas4 was only known to be associated with the CRISPR-Cas immune systems of organisms such as bacteria. However, the discovery of Cas4 solo proteins encoded also in bacteriophages that are one of bacteria's main predators, has added a new twist to the functional repertoire of the Cas4 family. The main objective of this study was to investigate vCas4 proteins with the specific aim of understanding their possible influence in the CRISPR-Cas system adaptation. The main questions answered in this thesis are:

1. Do vCas4 proteins interact with the Cas1-Cas2 complex of the CRISPR-Cas systems?
2. Do they have any influence on spacer acquisition?

3. If so, what's the mechanism underlying this influence?

Regarding the influence of vCas4 in the CRISPR system, in the case they do interact, two different hypotheses of how this interaction is established were addressed (Figure 4). The first is that the vCas4 interacts directly with the CRISPR system. In this case, the vCas4 binds to the Cas1-Cas2 complex inhibiting its ability to acquire new spacers even after infection. This way, the CRISPR system is not able to process an immune response against the foreign pathogen leading to bacterial death. We expected this hypothesis to be more prone to happen in systems in which Cas4 interaction with Cas1-Cas2 was already described since Cas4 can have a poisoning effect. The second hypothesis is that the vCas4 leads to the incorporation of wrong or non-functional protospacers by indirect interaction with the CRISPR-system. In this case, if vCas4 reveals nuclease activity, it can lead to the cleavage of the host DNA instead of the foreign invader nucleic acids. The cleaved host DNA is then integrated in the CRISPR array leading to auto-interference and bacterial death. This indirect interaction of vCas4 with the CRISPR system can also lead to the incorporation of wrong or non-functional protospacers (wrong PAM or inaccurate spacer size) that consequently allows successful phage infection and bacterial death.

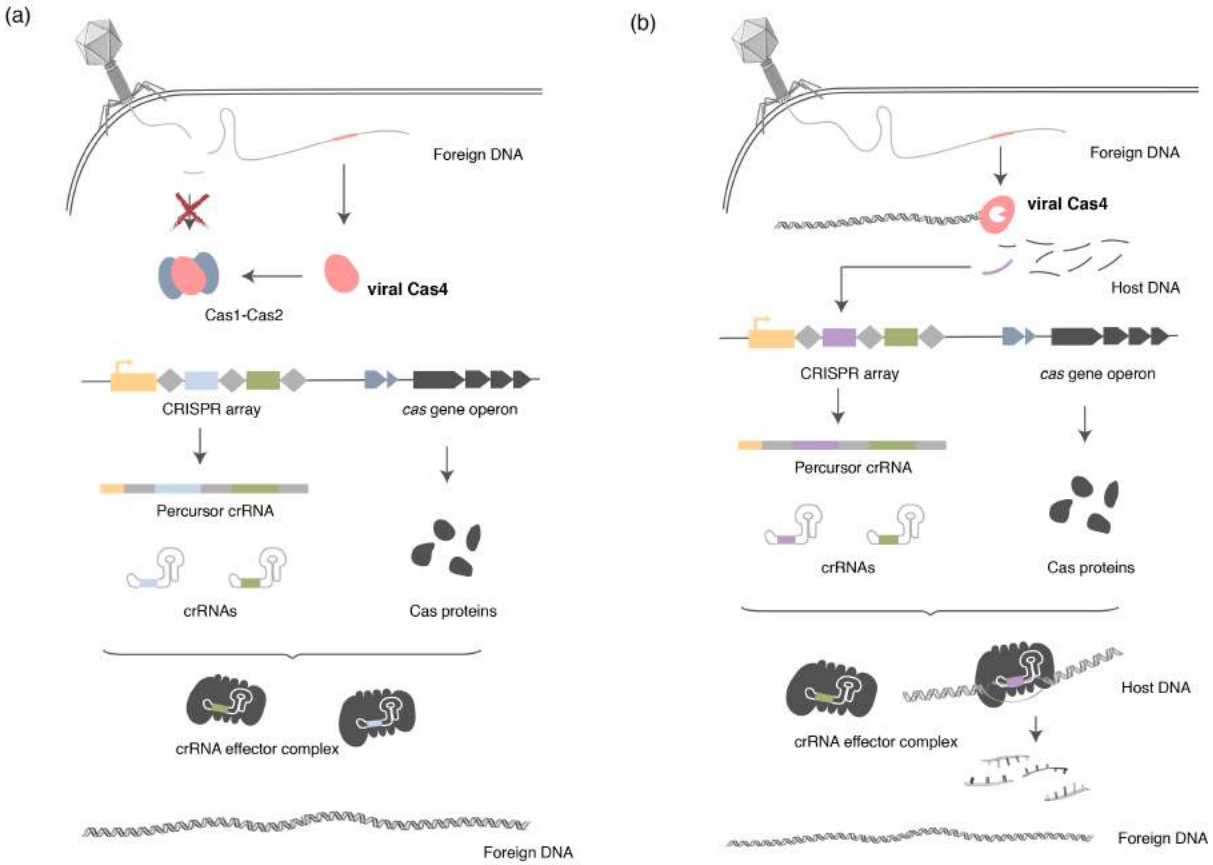

Figure 4: Hypothetical mechanisms of interaction between vCas4 and the CRISPR-Cas system. (a) The *vcas4* encoded in the foreign DNA genome is transcribed forming a viral Cas4 protein (in red) that interacts directly with the Cas1-Cas2 complex (in dark blue) enabling the acquisition of new spacers, target recognition and destruction of foreign DNA that is consequently kept intact inside the bacterial cell. (b) The *vcas4* encoded in the foreign DNA genome is transcribed forming a viral Cas4 protein with nuclease activity (in red) and leads to the destruction of the host DNA and consequent incorporation of these fragments as host derived spacers and consequent autoimmunity and destruction of the host DNA by the cr-RNA effector complex. The yellow fragment in the CRISPR array corresponds to the leader, the grey ones to the repeat sequences and the light blue and green fragments correspond to spacer sequences.

Chapter 2

Material and Methods

2.1 Bacterial strains

The bacterial strains used in this study were *E. coli* DH5α and *E. coli* BL21-AI. DH5α is an engineered strain of *E. coli* with high transformation efficiency mainly due to recA1 mutation that leads to high insert stability. Other mutations in this strain include Φ*80*Δ*dlacZM15*, Δ*(lacZYA-argF)U169*,*endA1*, *hsdR17(rK-mK+)*, *phoA*, *supE44*, *thi-1 gyrA96* and *relA1* [4]. BL21-AI is bacterial strain optimized for protein expression by deletion of *lon* and *OmpT* proteases leading to a reduction in degradation of heterologous protein expression. This strain carries also a chromosomal insertion of a cassette containing the *T7 RNA polymerase* (*T7RNAP*) gene allowing its expression under regulation of *araBAD* promoter [54].

2.2 Media and growth conditions

All bacterial cultures were grown in LB media (Luria-Bertani; 10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl) at 37℃ and continuous shaking at 180 rpm or in LBA plates (LB media supplemented with 15 g/L of agar), unless otherwise stated. When required, antibiotics and inducers were supplied to the final concentrations listed in Table 1 (see Table 3 for plasmids and corresponding selection markers).

Table 1: Antibiotics and inducers used in this study and corresponding final concentrations used.

	Compound	**Final concentration**
Antibiotics	Ampicillin	100 μg/mL
	Chloramphenicol	25 μg/mL
	Spectinomycin	50 μg/m
Inducers	L-arabinose	0,2% (w/v)
	IPTG	1 mM
	X-Gal	40 µg/mL

2.3 gBlocks

Some DNA fragments used in this study were chemically synthesized and ordered as gBlocks Gene Fragments from IDT - Integrated DNA Technologies. All the gBlocks used in this study are described in Table 2. The sequence of each fragment can be found in Appendix A.

Table 2: gBlock gene fragments used in this study.

gBlock	Description
LU11	Cas4 homolog gene encoded in LU11 *Pseudomonas* phage genome
KPP25	Cas4 homolog gene encoded in KPP25 *Pseudomonas* phage genome
CP30A	Cas4 homolog gene encoded in CP30A *Campylobacter* phage genome
C12LR	*Pseudomonas* type I-C Cas1-Cas2-Leader-Repeat
C4	*Pseudomonas* type I-C Cas4
C12IE	*Pseudomonas* type I-E Cas1-Cas2
LRIE	*Pseudomonas* type I-E Leader-Repeat

2.4 Plasmids

The plasmids used in this study and corresponding selection markers are described in Table 3.

Table 3: Plasmids used in this study.

Plasmid	Description	Resistance	Reference
p2AT	pET LIC cloning (2A-T)	Amp	Addgene # 29665
p13SS	pET His6 Sumo TEV cloning (13S-S)	Spec	Addgene # 48329
pACYC	pACYCDuet-1	Cm	Novagen # 71147
pCas12	pACYC with *E. coli* K-12 type I-E Cas1-Cas2	Cm	Not published
pTU223	p2AT with KPP25 vCas4	Amp	This study
pTU224	p2AT with LU11 vCas4	Amp	This study
pTU225	p2AT with CP30A vCas4	Amp	This study
pTU226	p13SS with KPP25 vCas4 (including His6 SUMO Tag)	Spec	This study
pTU227	p13SS with LU11 vCas4 (including His6 SUMO Tag)	Spec	This study
pTU228	p13SS with CP30A vCas4 (including His6 SUMO Tag)	Spec	This study
pTU229	p13SS with *P. aeruginosa* VA-134 type I-C Cas1-Cas2-Leader-Repeat (including His6 SUMO Tag)	Spec	This study
pTU230	pACYC with *P. aeruginosa* VA-134 type I-C Cas1-Cas2-Leader-Repeat	Cm	This study
pTU231	p13SS with *P. aeruginosas* VA-134 type I-C Cas4 (including His6 SUMO Tag)	Spec	This study
pTU232	p13SS with *P. aeruginosa* VA-134 type I-C Cas4 (deletion of His6 SUMO Tag)	Spec	This study
pTU233	p13SS with *P. aeruginosa* AZPAE14509 type I-E Cas1-Cas2 (including His6 SUMO Tag)	Spec	This study
pTU234	pACYC with *P. aeruginosa* AZPAE14509 type I-E Cas1-Cas2	Cm	This study
pTU235	p13SS with *P. aeruginosa* AZPAE14509 type I-E Leader-Repeat (including His6 SUMO Tag)	Spec	This study

Plasmids pTU223, pTU224 and pTU225 were obtained by Restriction Enzyme digestion and consequent ligation of both the p2AT plasmid and the gBlocks fragments. After construct confirmation, these fragments were amplified from each one of the plasmids and inserted in p13SS by Ligation-Independent Cloning, leading to the construction of pTU226-228. To construct the plasmids pTU229, pTU231 and pTU234, the GBlocks used were amplified and further inserted in p13SS by Ligation-Independent cloning. The Cas1-Cas2-Leader-Repeat fragment was amplified from pTU229 and inserted in pACYC by Restriction Enzyme Cloning (constructing pTU230). In both pTU229 and pTU230, an additional PCR-mediated deletion was preformed to remove an unwanted fragment downstream to *cas1* gene. pTU231 was obtained by LIC Cloning and by a PCR-mediated deletion, was further obtained the pTU232.

2.5 Polymerase Chain Reactions

To obtain the plasmids listed in Table 3 two different Polymerase Chain Reaction (PCR) were performed, either using Q5 DNA Polymerase (New England Biolabs) or OneTaq DNA Polymerase (OneTaq Hot Start Quick-Load 2X Master Mix with Standard Buffer from New England Biolabs). The components used in the mixture and their final volumes are indicated in Tables 4 and 5 for Q5 and OneTaq reactions, respectively. The PCR program applied was also dependent on the polymerase (see Table 6). The Annealing temperature (Ta) is dependent on the primers in use and corresponds to the rounded down average of the Melting temperature (Tm) of each individual primer. All the primers used in this study were purchased from IDT and can be found in Appendix B.

Q5 DNA Polymerase was used predominantly for amplifications of DNA in which a high nucleotide accuracy was required. OneTaq DNA Polymerase was commonly used to check the size of a DNA fragment in which the accuracy of the sequence was less relevant.

Table 4: Q5 DNA Polymerase mixture components and respective volumes (final volume of 50 µL) .

Component	Volume
Q5 Reaction Buffer (5x)	10,00 µL
dNTPS (10mM)	2,00 µL
Fw primer (10µM)	2,00 µL
Rv primer (10µM)	2,00 µL
Q5 DNA polymerase	0,50 µL
Template DNA (Genomic: 1 ng-1 µg; Plasmid:1 pg-1ng)	$\sim$ 1,00 µL
Milli-Q Water	to 50,0 µL

Table 5: One Taq DNA Polymerase mixture components and respective volumes (final volume of 20 µL) .

Component	Volume
OneTaq Master Mix with Standard Buffer (2x)	10,00 µL
Fw primer (10µM)	1,00 µL
Rv primer (10µM)	1,00 µL
Template DNA (Genomic: 1 ng-1 µg; Plasmid:1 pg-1ng)	$\sim$ 1,00 µL
Milli-Q Water	to 10,00 µL

2.6 Restriction Enzyme Cloning

Some of the plasmids used in this study were obtained by Restriction Enzyme Cloning in which the DNA fragments are digested using appropriate restriction enzymes and subsequently ligated.

For digestion, 3 µL of CutSmart buffer 10x (New Englands Biolabs Inc.), 1 µL of each enzyme, 1 µg of product to digest and the volume of Milli-Q water necessary to obtain the final volume of 30µL were incubated at 4 ℃ overnight. After that, restriction enzymes were inactivated by heat or by DNA purification (see DNA Purification). The restriction enzymes used in this study were EcoRI-HF, BamHI-HF, HindIII-HF, PstI-HF and KpnI-HF, all purchased from New England Biolabs, Inc..

Ligations of the digested fragments were performed in 20 µL. The amount of each fragment needed to obtain the final molar ration of 1:3 (vector:insert) was mixed with 2 µL of T4 Ligase buffer 10x, 1 µL of T4

Table 6: Q5 and OneTaq DNA polymerase PCR programs.

Step 1 (1 cycle)				
	Temperature (℃)		Time	
	Q5 PCR	One Taq PCR	Q5 PCR	One Taq PCR
Initial denaturation	95	95	2 min	10 min
Step 2 (20-30 cycles)				
	Temperature (℃)		Time	
	Q5 PCR	One Taq PCR	Q5 PCR	One Taq PCR
Denaturation	95	95	30 sec	30 sec
Annealing	Ta	Ta	30 sec	30 sec
Elongation	72	68	30 sec/kb	1 min/kb
Step 3 (1 cycle)				
	Temperature (℃)		Time	
	Q5 PCR	One Taq PCR	Q5 PCR	One Taq PCR
Final elongation	72	68	2min	2 min

ligase and Milli-Q water until final volume and this mixture was incubated at 4 ℃ overnight. This product was further transformed in *E. coli* DH5α cells (see Transformation via Heat Shock) and the final construct was confirmed by sequencing (see Sequencing).

2.7 Ligation-independent Cloning (LIC)

All the plasmids in which the His6 SUMO Tag was inserted were obtained by Ligation-Independent Cloning (LIC). In this cloning process, the fragments to insert were amplified by Q5 DNA polymerase PCR (see Polymerase Chain Reactions) and the vector (p13SS) was linearized using SspI restriction enzyme (500 ng of p13SS were incubated at 4 ℃ overnight with 1µL of restriction enzyme and Milli-Q in 10 µL of final volume). After purification of both PCR amplicon and linearized vector, LIC reactions were prepared as described in Table 7 and incubated at 22 ℃ for 30 minutes then 75 ℃ for 20 minutes. Finally, 3 µL of both LICed PCR and LICed vector were combined and incubated at room temperature during 10 minutes. This product was further transformed in *E. coli* DH5α cells (see Transformation via Heat Shock) and the final construct was confirmed by sequencing (see Sequencing).

Table 7: LIC reaction components and respective volumes (final volume of 20 µL) .

Component	Volume
Vector/Insert	$\sim$ 150 ng / $\sim$ 70-100 ng
dCTP or dGTP (25 mM stock)	2,00 µL
T4 DNA polymerase buffer (10x)	2,00 µL
DTT (100mM)	1,00 µL
T4 DNA polymerase	0,40 µL
Milli-Q Water	to 20,0 µL

2.8 PCR-mediated deletion

The PCR-mediated deletion process was implemented to remove DNA fragments from already constructed plasmids. This method was used to delete a mutation in plasmid pTU225 (three aditional thymine nucleotides present in CP30A *vcas4* gene), to remove an additional 80bp sequence from the gBlocks insertion in plasmids pTU229 and pTU230 (5'- ATGCGGCGACAGCTCAATACCCTATATGTCACCACCGAGGGCGCCTGGCTGAAGAAGGACGGAGCTAATGTCGTCATGGG - 3') and also to remove the fragment codifying for the His6 SUMO Tag in plasmid pTU232.

To complete the deletions, first, a Q5 DNA polymerase PCR using primers flanking the region to delete was performed (see Polymerase Chain Reaction). After DNA purification, 1 µL of dpnI (a methylation-sensitive restriction enzyme - New England Biolabs, Inc.) was added to the PCR product and this mixture was incubated at 4 ℃ overnight. This enzyme was inactivated by DNA purification and the 5' phosphorylation reaction was prepared as described in Table 8 by addition of T4 Polynucleotide Kinase (PNK - New England Biolabs, Inc.) in 20 µL final volume. After 45 minutes of incubation at 37 ℃, 1 µL of T4 DNA ligase was added and the final mixture was incubated at 4 ℃ overnight. This product was further transformed in *E. coli* DH5α cells (see Transformation via Heat Shock) and the final construct was confirmed by sequencing (see Sequencing).

Table 8: LIC reaction components and respective volumes (final volume of 20 µL) .

Component	Volume
5' phosphorylated DNA product	200 ng
ATP (10mM)	2,00 µL
T4 ligase buffer (10x)	2,00 µL
T4 Polynucleotide Kinase	1,00 µL
Milli-Q Water	to 20,0 µL

2.9 pGEM-T Vector System

The pGEM-T Vector System (Promega) was used to clone PCR products for White-Blue Screening. The ligation reactions were performed according to manufacturer indications [48] and the final product was further transformed in *E. coli* DH5α cells (see Transformation via Heat Shock).

2.10 Agarose Gel Electrophoresis

Visualization of DNA fragments was done using agarose gel electrophoresis. Gels were prepared mixing Agarose powder, LE, Analytical Grade (Promega Corporation) with 1x TAE Buffer (Promega Corporation) to final concentrations of 1-2 % agarose.

Gels were stained either using SYBR Safe (applied during gel preparation in a final concentration of 1x) or SYBR Gold (after electrophoresis, gels were incubated with agitation for 10-15min in 100 mL of TAE buffer (1x) and SYBR Gold in final concentration of 1%). SYBR Gold was used when high sensitivity to detect the nucleic acids was necessary. Both DNA gel stains were purchased from Invitrogen, Thermo Fisher Scientific. All agarose gels were imaged using Bio-Rad's ChemiDoc XRS+ System.

Except for PCR amplicons obtained using OneTaq mastermix, DNA Gel loading dye (Thermo Fisher Scientific) was added to the DNA samples to final concentration of 6x to allow visualization of DNA migration during electrophoresis. Eurogentec SmartLadder (200bp-10kb) and SmartLadderSF (100bp-1kb) were used as markers depending on the size of the expected fragments.

2.11 DNA Purification

Plasmid extraction from bacterial cultures was done using GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific). Purification of PCR products and DNA cleaning (as for restriction enzyme inactivation) was performed using GeneJET PCR Purification Kit (Thermo Fisher Scientific). Both kits were used as per manufacturer instructions.

2.12 Sequencing

DNA sequencing was done by Sanger method and outsourced to Macrogen Inc. Amsterdam. Fragments to sequence were sent along with the adequate primer in samples with final volume of 10 µL. The primer is the starting point of the sequencing being that the 100bp upstream to its binding site are not reliably sequenced.

All the samples were prepared combining 2 µL of primer(10mM) with the amount of template necessary to have around 250ng of PCR product in the final mixture.

2.13 Transformation

Transformation via Electroporation

Transformation of plasmids in *E. coli* BL21-AI was done via electroporation. Electrocompetent cells were prepared following a protocol adapted from Gonzales et al. (2013).

For transformation, a 100 µL aliquote of electrocompetent cells was incubated on ice with 100-200 ng of purified plasmid for few minutes. This mixture was then transfered to a chilled 2 mm electroporation cuvette (BioRad) and a pulse of 2500 V, 200 Ω and 25 µF was applied using an electroporator (BTX, Harvard apparatus). After electroporation, cells were immediately ressuspended in 1 mL of fresh LB medium and incubated at 37 ℃. After 1 hour of recuperation, 50 µL of cell culture were platted in LBA media supplemented with the antibiotic for which the plasmid to transform is resistant, allowing the selection of cells correctly transformed.

Transformation via Heat Shock

Transformation of plasmids in *E. coli* DH5α was done via heat shock.To prepare *E. coli* DH5α chemical competent cells, an independent colony from a culture growing overnight in LBA was inoculated in 50 mL of SOB media at 37 ℃ with agitation (180 rpm) to an OD600 of 0.3-0.6. From this culture, chemical competent cells were prepared using the *Mix & Go* E. coli Transformation Kit (Zymo Research) according to manufacturer instructions.

For transformation, the *E. coli* DH5α chemical competent cells were incubated on ice with 100-200 ng of purified plasmid for 30 minutes and thereafter heated to 42 ºC for 2 minutes. Cells were then incubated on ice for 2 minutes and immediately ressuspended in 1 mL of fresh LB medium and incubated at 37 ºC. After 1 hour of recuperation, 100 µL of cell culture and all the remaining pellet after centrifugation were plated in LBA media supplemented with the adequate antibiotic for selection of transformants.

2.14 *In vivo* spacer acquisition assays

To detect acquisition in both types I-E and I-C CRISPR systems of *P. aeruginosa* were transformed the Cas1-Cas2 complex and the vCas4 in *E. coli* BL21-AI. Besides that were also transformed the CRISPR array of the types I-E and I-C CRISPR systems of *P. aeruginosa* (a fragment containing the *leader* and *repeat* genes). In the case of CRISPR-Cas type I-C of *P. aeruginosa* and since this system already has Cas4 in its constitution, the acquisition was studied in the presence or absence of this native Cas4 (Figure 5).

From the transformed cells plated in LBA, single colonies were grown for 2 hours until OD600 of 0.1-0.3. After reaching this optical density, cells were induced by supplementation of L-arabinose and IPTG (see Table 1) and grown at 37 ºC and continuous shaking at 180 rpm overnight. In the case of the assays performed in type I-C, the cultures were additionally passed and induced again after 24 and 48 hours.

Spacer acquisition was monitored by PCR. To perform the reaction, 200 µL of each culture were, first, centrifugated and ressuspended in 50 µL of Milli-Q water, for salt removal. Then, 2 µL of cell suspension were used as template of the PCR reaction that was prepared accordingly to the type of CRISPR system in analysis. In the case of both *P. aeruginosa* CRISPR systems, the primers used were a forward primer annealing in the 3' end of the CRISPR repeat but mismatching the first nucleotide of spacer 1 (degenerate primer mix) and a reverse primer annealing in the vector backbone.

To allow the quantification of the results, the intensities of the non expanded and expanded bands were measured using the Image Lab™ Software report tool. With the values obtained was further quantified the normalized percentage of expanded band by dividing the obtained intensity by the sum of the intensities of both bands. This analysis was further complemented with an analysis of variance (ANOVA) to assess if the differences verified in the percentage of expanded band in the presence of vCas were significant or not when compared with the case vCas4 was not supplemented to the system (these results can be found in Appendix D).

The assays to detect acquisition were followed by the separation of the PCR products corresponding to the expanded CRISPR array from the parental ones. To perform this separation, the BluePippin (Sage Science) automated agarose-electrophoresis system was used (3 % agarose gel cassetes with Q3 Marker) as per manufacturer instructions. This system of separating DNA by size allows the selection of only expanded amplicons by selecting a range of separation that includes the size of the expanded CRISPR array, excluding the size of the parental one. The selected ranges were 190-210, 250-350 and 200-250 for type I-E and I-C CRISPR system of *Pseudomonas*. In the case of both *P. aeruginosa* CRISPR systems, the PCRs performed after automated gel extraction was performed using the same forward degenerate primer mix but with a different reverse primer that, in this case, matches spacer 1.

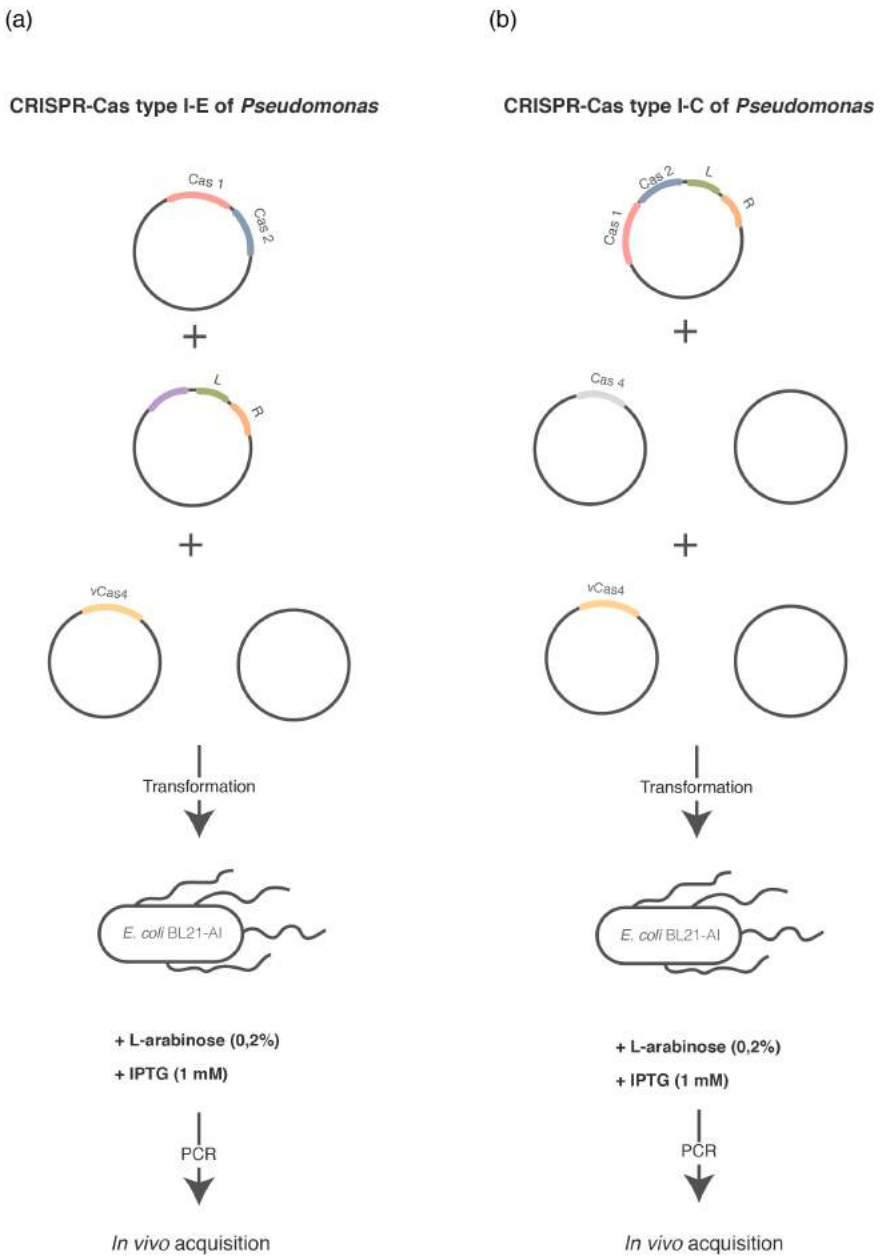

Figure 5: *In vivo* spacer acquisition assays. (a) Assay performed in CRISPR-Cas type I-E of *P. aeruginosa*. The plasmid with *cas1* and *cas2* genes and the plasmid with the *leader* (L) and *repeat* (R) genes along (His6-SUMO Tag in purple) were both co-transformed with the plasmids codifying the *vcas4* genes (KPP25, LU11 and CP30A) or, as a positive control, the empty plasmid (2AT) in *E. coli* BL21-AI cells. These cells were later induced by the supplementation of L-arabinose and IPTG to the medium until final concentrations of 0,2% and 1mM, respectively. After growth, a PCR was performed in order to evaluate the *in vivo* acquisition. (b) Assay performed in CRISPR-Cas type I-C of *P. aeruginosa*. The plasmid with the *cas1*, *cas2*, *leader* (L) and *repeat* (R) genes and either the plasmid carrying the Cas4 from type I-C of *Pseudomonas* or the empty plasmid (p13SS) were co-transformed with each one of the plasmids carrying the *vcas4* genes (KPP25, LU11 and CP30A) or, as a positive control, the empty plasmid (2AT) in *E. coli* BL21-AI cells. These cells were later induced by the supplementation of L-arabinose and IPTG to the medium until final concentrations of 0,2% and 1mM, respectively. After growth, a PCR was performed in order to evaluate the *in vivo* acquisition.

The PCR reactions (before and after BluePippin) were preformed in final volume of 50 µL using 2x OneTaq DNA polymerase (see Polymerase Chain Reactions). The PCR programs used were similar to the OneTaq DNA polymerase PCR described in Table 6, differing in the annealing temperature and number of cycles.

The *in vivo* acquisition PCRs were done with annealing temperatures between 58 and 60ºC and 30 to 35 cycles and, after BluePippin, Ta was 62ºC and the PCR was performed with 25 cycles. The visualization of amplicons was done in 2 % agarose gels, loading 10 µL of each sample.

Expanded CRISPR array sequencing and protospacer analysis

The expanded CRISPR array collected from BluePippin was inserted in the pGEM-T vector (as explained in pGEM-T Vector System) and transformed in *E. coli* DH5α cells for Blue-White Screening. Thereby, the white colonies (in which the new acquired spacers were inserted) were grown in a 96-well plates containing LBA and adequate antibiotics. The sequencing of the plated colonies was done by Sanger method and outsourced to GATC (Eurofins Genomics).

The sequencing results were further analysed and the protospacers acquired in the expanded CRISPR array were identified. This protospacer were analyzed by a Basic Local Alignment Search Tool (BLAST) search against the *E. coli* BL21-AI genome or the inserted plasmids sequence, according to the CRISPR system in analysis. Finally, the upstream sequence of each protospacer was analyzed with Weblogo to determine the PAM consensus sequence.

2.15 Protein Purification

During this study, different proteins were overexpressed and purified with two different final objectives: to further perform *in vitro* protein activity assays or to infer about co-purification and, consequently, protein-protein interactions. That way, different protein purification workflows were followed (Figure 22).

With the objective of determining protein activity, the protein in study were subjected to an additional protein purification by Size-Exclusion Chromatography (SEC) using the ÄKTA pure system.

On the other hand, with the objective of purifying a protein to conclude about its interaction with other specific proteins (to evaluate co-purification), the Ni-NTA affinity chromatography was followed by Mass Spectrometry analysis.

2.15.1 Protein expression and purification by Ni-NTA Affinity Chromatography

To overxpress the protein of interest, plasmids in which this protein was codified along with the polyhistidine sequence were transformed in BL21-AI cells. From a liquid culture of this cells growing overnight at 37ºC 180 rpm, 2L of LB media were inoculated and grown at 37ºC, 180 rpm to OD600 of 0.3-0.6. When reaching this optical density, the culture was kept on ice during 10 minutes, induced by supplementation of L-arabinose and IPTG and grown at 20 ℃ and continuous shaking at 180 rpm overnight. Cells were then harvested by centrifugation at 6000 rpm and 4 ℃ for 10 minutes, ressuspended in 50 mL of chilled Lysis Buffer and cOmplete EDTA-free Protease Inhibitor Cocktail (Sigma-Aldrich) and, later, lysed by French Press (1000 bar). The lysate was cleared by centrifugation for 30 min at 16000 rpm and 4ºC and filtered through a 0,45mm syringe filter. 500 µL of HIS-Select® Nickel Affinity Gel (Sigma Aldrich) were washed with chilled Lysis Buffer and incubated with the clarified lysate at 4 ℃ for 30 minutes. The icubated lysate and resin were then load in a gravity disposable column and the first fraction was collected ('Before Wash' sample). The column was washed two times with Wash Buffer ('Wash' samples) and bound proteins were eluted in Elution

Buffer (5 elutions of 500µL each were collected). The composition of all the buffers used can be found in Table 9 (the final pH of all buffers was adjusted to 7,5).

After protein purification, the protein concentration and A280 of each elution was measured using NanoDrop (Thermo Fisher Scientific). The visualization of the purified proteins was done by SDS-page eletrophoresis (see Polyacrilamide Gel Eletrophoresis) and samples were stored at -80ºC.

Table 9: Components of all the buffers used for protein purification and respective concentration.

	Lysis Buffer	**Wash Buffer**	**Elution Buffer**
HEPES	50 mM	50 mM	50 mM
KCl	300 mM	300 mM	300 mM
Glycerol	5%	5%	5%
Imidazol	5mM	20mM	300mM
DTT	1mM	1mM	1mM
Triton X-100	0,1%	-	-

2.15.2 Size-Exclusion Chromatography (ÄKTA Pure system)

To perform the size-exclusion chromatography, the elutions collected from the previous chromatography procedure were pooled together and concentrated to a final volume of approximately 600 µL using a centrifugal filter unit (EMD Millipore Amicon Ultra from Thermo Fisher Scientific). After centrifugation at 4℃ during 10 minutes, 4800 rpm, the supernatant was applied to a Superdex 200 10/300 GL column connected to an ÄKTA purifier system (both purchased from GE Healthcare) previously washed with fitered Milli-Q and Elution Buffer (see Table 9). The sample was eluted with Elution Buffer at a flow rate of 0.3 mL/min at 4 ℃ and detected at 260 and 400 nm. Fractions of 1 mL were collected.

Fractions of interest were collected and its protein concentration and A280 was measured using NanoDrop (Thermo Fisher Scientific). The visualization of the purified proteins was done by SDS-page eletrophoresis (see Polyacrilamide Gel Eletrophoresis) and samples were stored at -80ºC.

2.15.3 Polyacrilamide Gel Eletrophoresis

Visualization of the purified proteins was done using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The gels used were purchased from Bio-Rad (4–20% Mini-PROTEAN TGX PRecast Protein Gels). To prepare the samples, 20µL of each elution were mixed with 6X SDS Protein Loading Buffer (Laemmli buffer; 375 mM Tris.HCl, 9% SDS, 50% Glycerol and 0.03% Bromophenol blue) and incubated at 95ºC during 10 minutes, to allow protein denaturation. Samples were harvested by centrifugation at 5 rpm during 4 minutes and 20 µL of supernatant were loaded in the SDS-PAGE gel. The gel was covered in 1x TGS Buffer (25mM Tris, 192mM glycine, and 0.1% SDS) and PageRule Prestained Protein Ladder from Thermo Fisher Scientific (10-180 kDa) was used as marker.

After eletrophoresis, the gel was incubated with InstantBlue Comassie Protein Stain (from Expedeon) during 15 minutes and imaged using Bio-Rad's ChemiDoc XRS+ System.

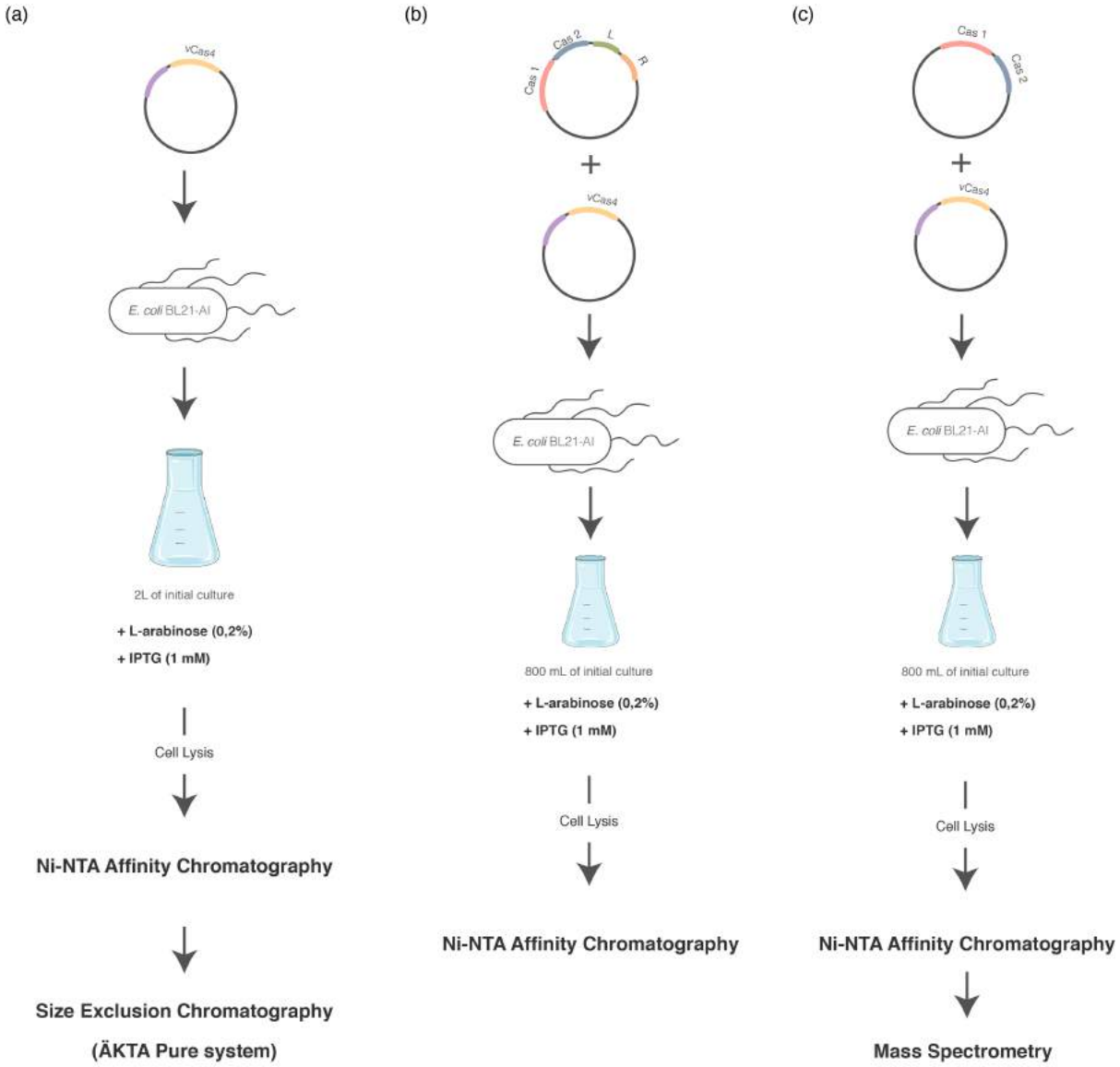

Figure 6: Protein purification workflows. (a) Each one of the plasmids carrying the *vCas4* genes (KP225, LU11 and CP30A) including the His6-SUMO Tag (in purple) were transformed in *E. coli* BL21-AI that were further incubated in 2L of initial culture and then induced with L-arabinose and IPTG. After growth, cells were lysed and vCas4 proteins were purified by Ni-NTA Affinity Chromatography followed by Size-Exclusion Chromatography (using the ÄKTA Pure systems) (b) Co-purification assay in CRISPR-Cas type I-C of *P. aeruginosa*. The plasmid with the *cas1*, *cas2*, *leader* (L) and *repeat* (R) genes was transformed with the each one of the plasmids carrying the *vCas4* genes (KP225, LU11 and CP30A) including the His6-SUMO Tag (in purple) were transformed in *E. coli* BL21-AI and then induced with L-arabinose and IPTG. After growth, cells were lysed and vCas4 proteins were purified by Ni-NTA Affinity Chromatography. (c) Co-purification assay in CRISPR-Cas type I-E of *P. aeruginosa*. The plasmid with the *cas1* and *cas2* genes was transformed with the each one of the plasmids carrying the *vCas4* genes (KP225, LU11 and CP30A) including the His6-SUMO Tag (in purple) were transformed in *E. coli* BL21-AI and then induced with L-arabinose and IPTG. After growth, cells were lysed and vCas4 proteins were purified by Ni-NTA Affinity Chromatography and further subjected to an additional Mass Spectrometry analysis.

2.16 Mass Spectrometry

Samples subjected to Mass Spectrometry were prepared and outsourced to Bokinsky Lab (Bionanoscience Department, TUDelft).

For sample preparation, 50 µg of protein in solution were mixed with 50mM NH_4HCO_3 to a final volume

of 93,5 µL and 1 µL of 0,5 M DTT and incubated for 1 hour at room temperature. Then, after adding 2,7 µL of iodoacetamide (0,55M), the mixture was incubated in dark for 15 minutes room temperature. 1 µL of ProteaseMAX surfactant (1%) and 1,8 µL of trypsin (1 µg/µL) were added and the mixture was incubated at 37℃ overnight. Samples were then centrifugated at 14000 x *g* for 10 seconds and 5µL of trifluoracetic acid were supplemented. Ttrypsin was inactivated by the incubation of this mixture for 5 minutes at room temperature. Samples were finally harvested by centrifugation at 14000 x *g* for 10 minutes and stored at -20 ℃ until LM-MS analysis.

2.17 *In vitro* Nuclease Activity Assays

After purification of vCas4 by Ni-NTA affinity chromatography followed by size exclusion chromatography, assays to determine the *in vitro* nuclease activity of these proteins were performed. The activity of vCas4 was tested in circular and linear double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA) using circular and linearized pACYC plasmid and M13 DNA (M13mp18 ssDNA purchased from New Englad Biolabs), respectively.

The *in vitro* nuclease activity assays were performed only for KPP25 and CP30A vCas4 proteins. 500 nmol of the purified protein were mixed with 100 ng of DNA substrate in 10x reaction buffer (20 mM MES pH 6.0, 10 mM DTT, 100 mM potassium glutamate„ supplemented with 10 mM $MgCl_2$ or 10 mM $MnCl_2$ as indicated) and Milli-Q water until final volume of 10 µL. In the case of KPP25 vCas4, this mixture was incubated at 37ºC during 2 hours. With samples on ice, the reaction was quenched by the addition of Proteinase K (New England Biolabs) to a final concentration of 10 mM and incubation at 37 ℃ for 45 minutes. To evaluate the activity, samples were mixed with 6X Loading dye, visualized in a 1% agarose gel and stained with SYBR Gold (see Agarose Gel Eletrophoresis).

In the case of CP30A vCas4, the mixture of purified protein, DNA substrate and buffer was incubated for 0, 5, 10, 30 or 60 minutes. To stop the reaction, along with the Proteinase K, EDTA was also supplemented to a final concentration of 10 mM and the mixture was incubated for 45 minutes at 37 ℃. Visualization of the DNA fragments was processed as described above.

2.18 Exonuclease Activity Assays

Supplementar exonuclease activity assays were performed using purified CP30A vCas4 protein. In this study, 50 nM of protein were mixed with 5 nM of DNA substrates (chemically synthesised oligonucleotided incorporating a fluorescent label at the 3' or 5' ending purchased from Integrated DNA Technologies).The 10x reaction buffers (with 10 mM $MgCl_2$ or 10 mM $MnCl_2$ as indicated) were supplemented and reactions were performed as described above for the case of CP30A vCas4 activity (see *In vitro* Nuclease Activity Assay). To stop the reaction, formadide loading mix was added to the mixture (1:1) and heated to 95℃ during 10 minutes. Samples were further separated on a PAGE-denaturing gel (20% polyacrylamide, 7M urea, 1× TBE).

2.19 Bioinformatic Analysis

All the bioinformatic analysis performed during this study were done using Geneious 9.1.8. and, depending on the analysis, supplementary plugins were used. For multiple sequence alignment was used the MAFFT plugin [13] and to find CRISPR locus the CRT plugin (CRISPR Recognition Tool) [**?**].

The phylogenomic tree of vCas4 was obtained using the RAxML plugin after MAFFT alignment of all the aminoacid sequences of the vCas4 proteins that were previously found by BLAST analysis (using NCBI). The RAxML (Randomized Axelerated Maximum Likelihood) is an implementation of maximum-likelihood (ML) phylogeny estimation that operates on protein sequence alignments [**?**].

Chapter 3

Results

Three different approaches were carried in the study of vCas4 proteins: bioinformatic analysis, *in vivo* acquisition assays and biochemistry studies. The bioinformatic analysis allowed the study of the vCas4 phylogenomics and establishment of three different vCas4 proteins to perform further experimental assays. With the *in vivo* acquisition assays was possible to study the vCas4 influence in the acquisition of new spacers by the CRISPR system. Were determined the rates of novel spacers acquired in the presence or absence of vCas4 and also the origin, length and consensus PAM of this newly acquired spacers. Additionally, were performed co-purification assays that allowed the study of protein-protein interactions between vCas4 and the Cas1-Cas2 complex. With the final biochemistry analysis was possible to determine the activity of vCas4 proteins. The results obtained are presented below.

3.1 Bioinformatic Analysis

The first objective of this study was to compile the highest amount possible of genes encoding vCas4 and constitute a database of these proteins. To compile them, the sequences encoding Cas4 proteins found in phage genomes that were already known from the phylogenomic study of the Cas4 nuclease family, performed by Hudaiberdiev *et al.*, 2017, were analyzed by the NCBI BLAST tool to identify possible similar genes located in the genomes of different phages. From this analysis were retrieved 134 sequences in total.

These genes were then translated into protein sequences and aligned with all the non-redundant protein sequences in the NCBI database using the PSI-BLAST tool. With this analysis was possible to verify if the compiled proteins belonged to the Cas4I-AI-BI-CI-DII-B superfamily by the evaluation of domains conserved by the hypothetical vCas4 proteins and the ones known to be part of this superfamily. From this analysis were retrieved 112 proteins which domains were conserved. This way was possible to obtain a database of vCas4 proteins with a total amount of 112 homologs.

These vCas4 homologs were found to be encoded in different types of phages such as *Mycobacterium, Vibrio, Salmonella, Xanthomonas, Campylobacter, Pseudomonas, Bacillus, Gordonia, Brevibacillus, Escherichia, Ralstonia, Erwinia, Roseobacter, Acinetobacter, Pseudoalteromonas, uncultured Mediterran, Croceibacter, Achromobacter and Shewanella* phages and also in some virus such as the Acidianus filamentous, Halovirus, Sulfolobus and the EBPR siphovirus showing evidence that it is possible to find vCas4 proteins encoded in a high diversity of viral organisms. Additionally, it was also possible to verify that a big amount of them were encoded in *Mycobacterium* phage genomes (40 out of 112 proteins). These Mycobacterium

bacteriophages were under the scope of a wide phage study [47] that resulted in the sequencing of several of this phages. This way, since a lot of similar phage genomes were available, several vCas4 proteins were found in their genomes.

With the obtained database were further performed phylogenomic studies in order to evaluate the possible similarities between these proteins. To perform these assays, the 112 vCas4 proteins were aligned by MAFFT alignment [13] being also included the Cas4 proteins known to be associated with type I CRISPR-Cas systems (from types I-A, I-C, I-D and the Cas4 domain of the type I-U fusion) and an additional protein encoded in *Thermoproteus tenax virus* (TTV1). In this virus, the Cas4 gene is split in two, with the N-terminal portion becoming a structural coat protein (TP1) [25]. This resulted in the inactivation of the nuclease activity of the Cas4 protein by the lost of some of the catalytic amino acid residues [25]. This way, since it is known that this protein has lost some residues when compared to the remaining Cas4 proteins, it was expected to detect which residues were conserved by analysis of the protein alignment. This protein was then included in the alignment because it might allow the identification of proteins with probability of showing similar structural function by the identification of proteins missing the same identified residues.

From the alignment it is possible to see that 4 cysteine residues (3 in the C-terminal and 1 in the N-terminal) are very well conserved in all the vCas4 proteins in study, including the TTV1 protein and the Cas4 proteins encoded in the CRISPR loci (Figure 7). This four cysteines are known to be highly conserved in Cas4 proteins which are presumably responsible for the coordination of the iron-sulfur cluster [72]. The fact they are conserved in almost all the proteins presented in the alignment shows high evidence that these proteins might have similar function. By the evaluation of this residue, it is not possible to isolate the TTV1 protein from the remaining Cas4 proteins since the four cysteines are also conserved in its sequence.

However, structural studies of Cas4 have shown that not only the 4 cysteines are conserved but also some other residues [33]. These residues constitute the RecB motif and they are known to be highly conserved in the Cas4 nucleases family [72]. In this residues is included a lysine known to be responsible for the DNA nuclease activity of the Cas4 proteins [30] and an additional aspartic acid residue [72] also relevant in the nuclease activity demonstrated by Cas4. It is possible to see that, as the four cysteine residues, the aspartic acid residue (D) is very well conserved in all the proteins included in the alignment (Figure 7). In the case of the lysine residue it is possible to see that, in fact, it is very well conserved in the majority of the Cas4 proteins presented. However, curiously, it is possible to see an exception: the TTV1 Cas4 protein. Instead of a lysine residue this protein as an arginine in that position. The fact that one residue involved in the nuclease activity changed might be the reason why this protein evolved to be part of the nucleocapsid structure of phages. Since this is the only important residue that is not conserved in the TTV1 protein when compared to the remaining Cas4 homolog proteins, we can conclude that, besides that fact this protein corresponds to a division of a Cas4 protein, the fact this residue was lost might also have relevance in the fact that this protein shows a different function when compared to the already known CRISPR associated Cas4 proteins.

Even if the cysteine residues are very well conserved in the majority of the vCas4 proteins, there are some cases in which only the lysine and aspartic acid residues are conserved such as the case of *Escherichia* phage PBECO4 or the *S. monocaudivirus* SMV4. It could also be interesting to understand what the vCas4 proteins maintain their activity in the case the cysteine residues are missing but the RecB motif is unaltered.

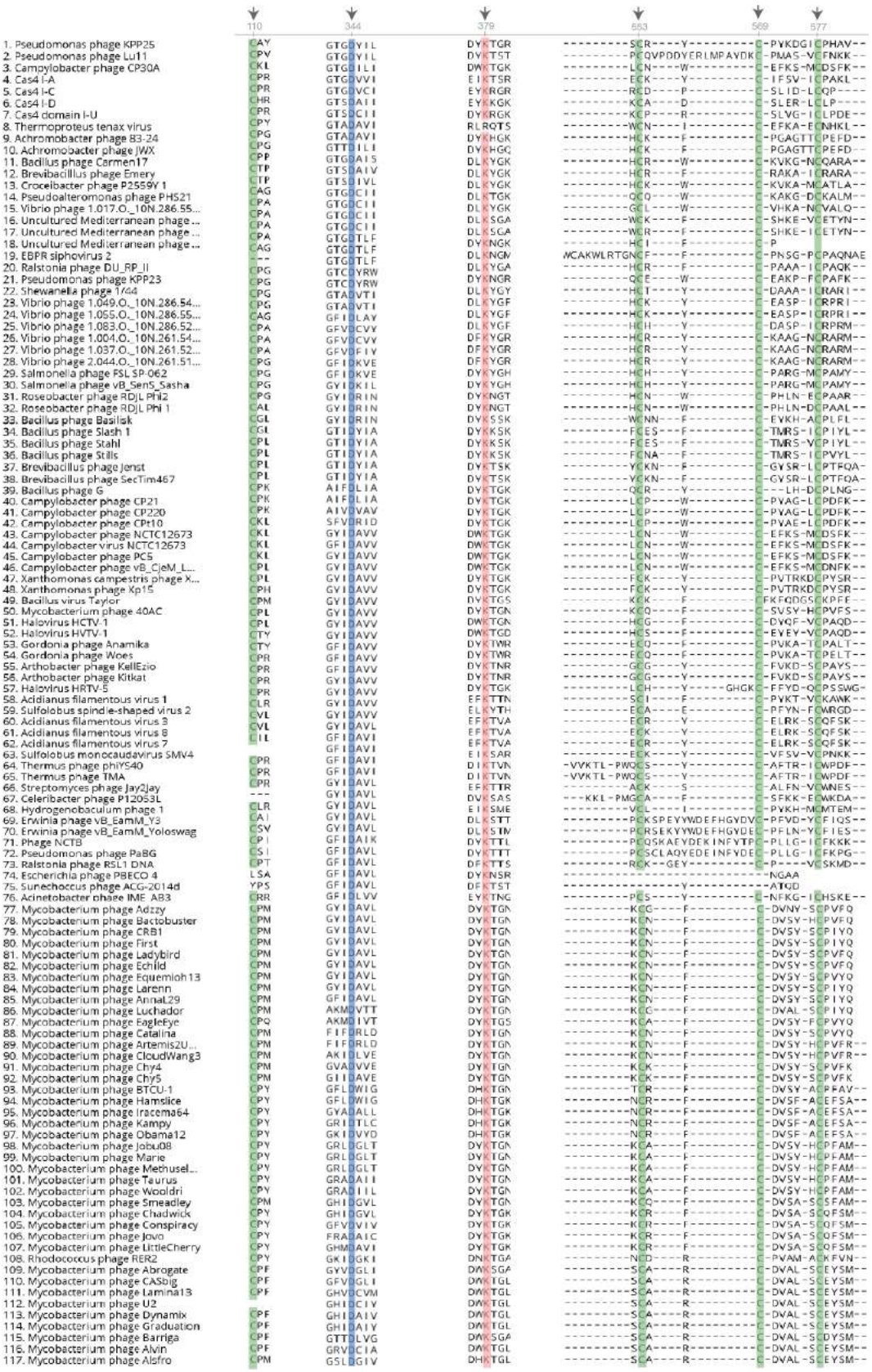

Figure 7: Protein alignment of vCas4. In this analysis were included of the 112 vCas4 proteins belonging to the established database, the five Cas4 proteins known to be associated with the types I-A, I-C, I-D and the Cas4 domain of the type I-U fusion and a protein encoded in *Thermoproteus tenax virus* (TTV1). Obtained by MAFFT alignment. Highlighted are the four conserved cysteines (green), the conserved aspartic acid (blue) and the conserved lysine (red).

This analysis allow us to conclude that the vCas4 proteins have high similarity to Cas4 proteins associated with type I CRISPR-Cas systems and also that with the analysis of the domains conserved in the aligned proteins it is possible to spot proteins that, as the one encoded in TTV1, show differences in their function.

With the obtained results, was further studied the possibility of having the same data in a different conformation, maintaining the possibility of detecting proteins as TTV1. This way, with the given alignment was obtained a sequence similarity dendogram. With this analysis was studied the possibility of using phylogenomic trees as a complementary tool to distinguish between vCas4 proteins with DNA-related activity and proteins that, even if derivated from Cas4, show different function or, additionally, which vCas4 are more related to the Cas4 proteins known to be associated with the CRISPR-Cas systems. When compared to the protein alignment, the phylogenomic trees have the advantage of being easier to analyze and allowing the detection of evolutionary patterns and evolutionary relations between the proteins in analysis.

In the phylogenomic tree was obtained, as expected, a big central cluster of vCas4 proteins encoded in Mycobacterium phages (Figure 8). Since these proteins were all isolated from phages infecting the same organism [47], as expected, the vCas4 proteins encoded in their genomes have high similarity and, consequently, clustered together in the phylogenomic tree.

However, contrarily to supposed, the TTV1 protein clustered together with other vCas4 proteins and not independently. Contrarily to expected, in this case, if the vCas4 proteins were analyzed only by the phylogenomic tree obtained, this protein with lost functionality wouldn't be identified. It is important that the phylogenomic analysis allows the recognition of genes, that even if very similar to the other Cas4 proteins, have evolved to have a different function.

Moreover, the Cas4 proteins known to be associated with type I CRISPR-Cas systems (from types I-A, I-C, I-D and the Cas4 domain of the type I-U fusion) that were included in this study are not clustered together but placed in different regions of the phylogenomic tree contrarily to what was previously described in the phylogenomic study performed by Hudaiberdiev *et al.*, 2017. In that study, it was found that the majority of the viral Cas4 proteins would cluster together in independent groups of proteins even if it was also found vCas4 proteins included in clusters of CRISPR-related Cas4 proteins. However, this correlation is not possible to find in the phylogenomic tree presented. Even if it was possible to detect the similarities and differences between the Cas4 proteins from type I CRISPR-Cas systems and the other vCas4 proteins in the alignment, these properties are lost in the analysis of the phylogenomic tree. It might by related to the fact that the phylogenomic tree obtained in our study includes an amount of proteins sequenced considerably lower (117 proteins in total) to the amount of proteins included in the study performed by Koonin *et al.* in wich 7060 proteins were evaluated. Nevertheless, it allows us to conclude, once again, that the protein alignment is a better tool to assess about possible differences between the Cas4 protein homologs.

With this conclusion and since it was necessary to choose proteins to perform further experimental assays from the vCas4 proteins database, this selection was done based on the protein alignment instead of analyzing the phylogenomic tree. The first criteria of choice was the conservation of the four cysteines and the one lysine domains since it was shown that these domains are highly conserved in the Cas4 proteins but not in the TTV1 capsid protein. The proteins chosen were the ones encoded in *Campylobacter* phage CP30A, *Pseudomonas* phage KPP25 and *Pseudomonas* phage LU11.

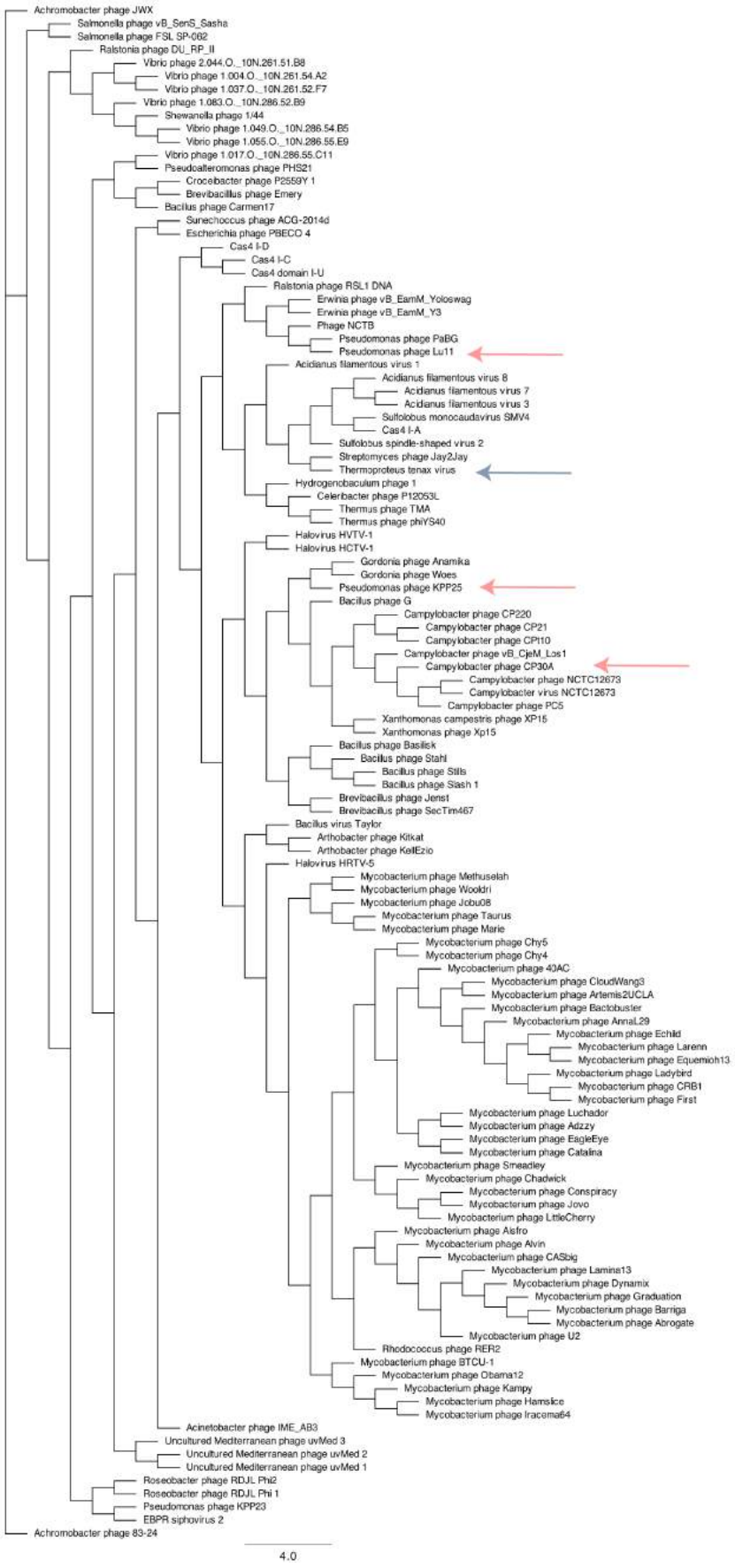

Figure 8: Phylogenomic tree of vCas4. Sequence similarity dendogram obtained from the alignment of the 112 vCas4 proteins including also the Cas4 proteins known to be associated with the type I CRISPR-Cas systems. In red are marked the vCas4 proteins in study and, in blue, the Cas4 homolog encoded in TTV1 genome.

The vCas4 protein encoded in CP30A (hereafter, vCas4 CP30A) was chosen since it was previously shown that this protein is responsible for stimulating the acquisition of host-derived spacers by the *Campylobacter* type II-C CRISPR-Cas systems (lacking cas4) and that, by an uncharacterized mechanism, the *Campylobacter* phage appears to use these vCas4 protein to escape from host immunity [16, 17]. This way, this protein was chosen as a control since it was already studied and also because with the evaluation of the effect of this protein in acquisition of other types of CRISPR-Cas systems we can conclude about the universality of its activity.

Moreover, were also chosen the vCas4 proteins encoded in the *Pseudomonas* phages KPP25 and LU11 (hereafter, vCas4 KPP25 and vCas4 LU11). These proteins were chosen since, as referred, they are encoded in *Pseudomonas* phages. This is an advantage since these bacteria (and the correspondent genes encoding the CRISPR-Cas system) are accessible and well studied. Since the main objective of this study is to detect the vCas4 influence in the CRISPR-Cas system mechanism, it is ideal to study proteins encoded in phages that infect bacteria possible to be tested *in vivo*.

Another advantage of these vCas4 proteins is the specific bacterial strains infected by the phages in which they are encoded. Even if both LU11 and KPP25 vCas4 are encoded in *Pseudomonas* phages, these phages differ in the bacterial strains they infect. The *Pseudomonas* phage LU11 infects specifically *Pseudomonas putida* and KPP25 infects *Pseudomonas aeruginosa*. No CRISPR-Cas system have ever been identified in *Pseudomonas putida*, however, in *Pseudomonas aeruginosa* three different CRISPR-Cas systems were identified [7]. This bacterial strain is known to have types I-E, I-C and I-F CRISPR systems. The study of a phage that infects a bacterial strain that possesses a type I-C CRISPR-Cas system (that includes a Cas4 protein in the CRISPR loci) is a big advantage since it allows the study of competition between the native Cas4 of the system and the vCas4. Moreover, the fact this phage infects type I-E CRISPR-Cas systems is also a big advantage since this system is also encoded in some *E. coli* strains being one of the most well studied systems.

Even if it is already known that the phylogenomic tree have to be carefully used in the prediction of genes function, it is possible to see that the proteins chosen to study are closely located to the TTV1 proteins in the obtained phylogenomic tree (Figure 8). Since it is important to make sure that the vCas4 proteins in study were not related with this protein, an additional analysis was performed. This way, was further analyzed the localization of the genes encoding for the vCas4 proteins in study in the phage genome.

This analysis was performed since it is known that the phage genomes can be divided in different clusters of genes [9] according to genes' function (comprising early, early-middle, middle and late clusters of genes). This way, by studying the localization in the phage genome of each one of the genes encoding the vCas4 proteins and annotation of the surrounding genes it was possible to determine in each cluster were located the vCas4 genes in study and further confirm either if the protein TTV1 would be integrated in a late cluster, as expected due to its known function, and also if the chosen proteins to study would be integrated in the early or early-middle clusters with high probability of showing DNA-related activity, as desired.

Usually, in the early region are comprised genes which function is not well defined. The early middle and middle region contains genes that are associated with DNA metabolism. In the early-middle cluster is possible to find proteins such as DNA-binding proteins, DNA primases, DNA helicases and DNA ligases

and, in the middle cluster proteins as DNA polymerase, exonuclease, endonuclease and RNA polymerase. In the late cluster are encoded structural proteins (as head-tail proteins, capsid proteins, holin and endolysin) [56, 71]. The composition of phage genomes in the described clusters is closely related to the phage life cycle (Figure 9 a). After infection, the phages express the proteins in the early and early-middle clusters and then, during expression of early-middle proteins, occurs phage DNA replication inside the pathogen. Only after expression of middle and late cluster proteins occurs effective cell lysis.

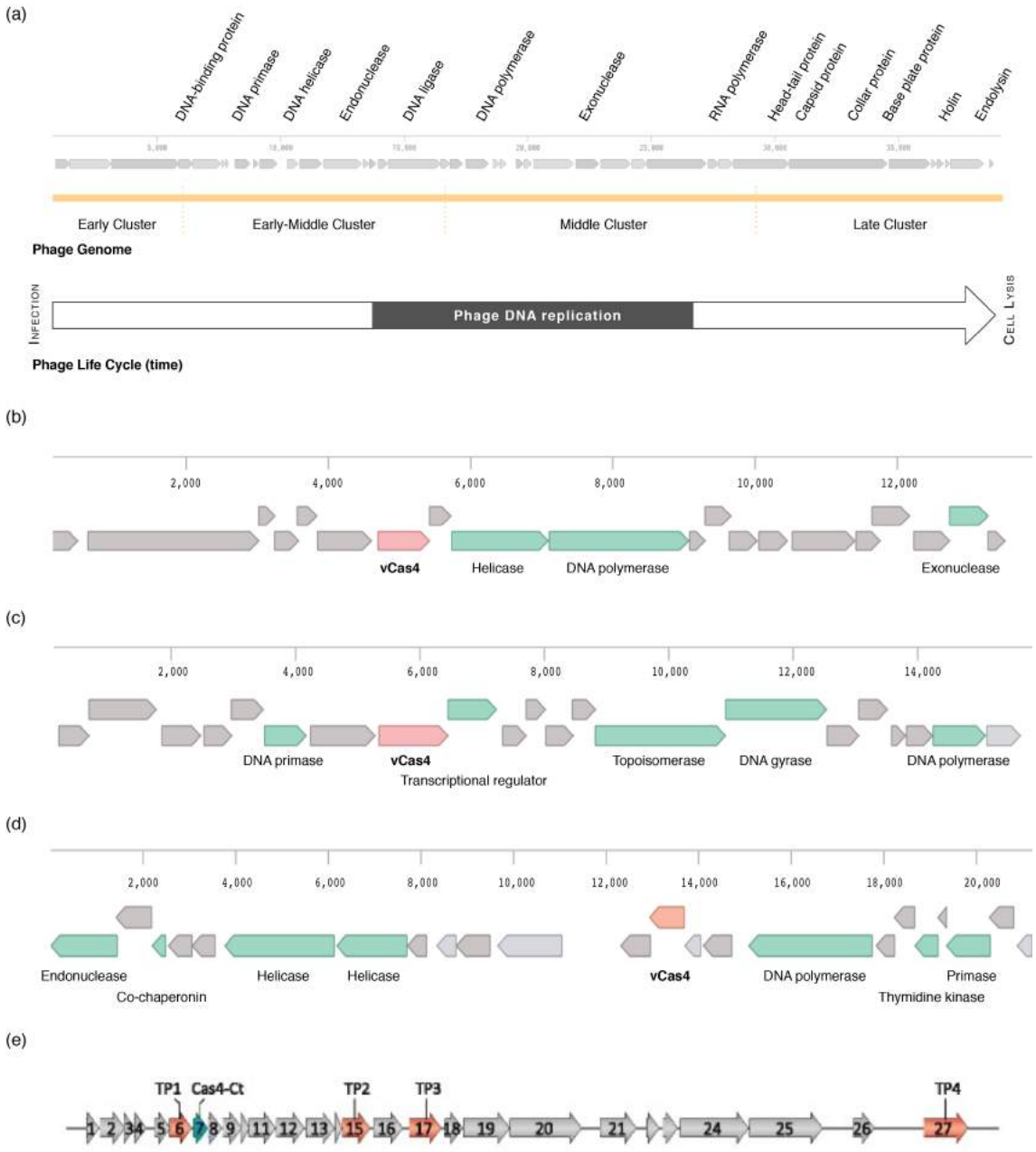

Figure 9: Localization of vCas4 proteins in the phage genome and phage life-cycle (a) Classification of the phage genome in clusters of genes depending on their function (listed are examples of genes encoded in each one of the clusters) and its correspondence to phage life-cycle (b) Annotation of genes in the proximity of KPP25 *vCas4* (in red). (c) Annotation of genes in the proximity of LU11 *vCas4* (in red). (d) Annotation of genes in the proximity of CP30A *vCas4* (in red). (e) Localization of the nucleocapsid protein derivated from the CRISPR-associated Cas4 nuclease encoded in Thermoproteus Tenax Virus (Cas4-Ct protein highlighted in green). The proteins TP1-TP4 correspond to coat proteins. From Krupovic *et al.*, 2015.

It was found that the KPP25 vCas4 is encoded downstream to a helicase and DNA polymerase proteins (Figure 9). The function of all the remaining proteins was not possible to determine since their analysis using the NCBI BLAST did not retrieve any similar proteins which function was already known. DNA helicase and polymerase are known to be part of the early-middle and middle clusters, respectively. Since this protein is located downstream to these two proteins we can conclude that it belongs to the early-middle cluster.

The gene encoding for this vCas4 LU11 is preceded by a DNA primase, one of the first genes being expressed by phages after infection. Besides that, it is followed by a trancriptional regulator, a topoisomerase, a DNA girase and DNA polymerase (Figure 9). All these proteins are located in the early-middle and middle clusters. Thus, it is possible to conclude that LU11 vCas4 is also located in the early-middle cluster.

The gene encoding CP30A vCas4 is preceded by a primase, thymidine kinase and DNA polymerase and followed by two helicase proteins, a co-chaperonin and an endonuclease (Figure 9). Following the same analysis as before, since all the genes encoding for these proteins are located in the early-middle and middle clusters and no structural proteins were found, we can conclude that CP30A vCas4 is located in the early-middle region of the phage genome.

Finally, was analyzed the localization of the TTV1 gene. As explained, due to the structural properties of this Cas4 homolog gene, it would be expected to find it in the late cluster of the phage genome. This is confirmed since the genes TP1-TP4 encode for capsid proteins (Figure 9). With this evaluation it was then possible to conclude that, contrarily to the TTV1, as the other genes known to be part of the early-middle cluster, the vCas4 proteins chosen to be studied might have function and activity related to the metabolism of nucleic acids.

Finally, it is possible to conclude that with the objective of identifying proteins from the database of vCas4 proteins that, as TTV1 are related to Cas4 but evolved to have a different function analysis, the protein alignment have to be analyzed instead of the phylogenomic tree and, the genes encoding for these proteins, might be supplementary localized in the phage genome and phage life cycle. Even if the phylogenomic tree allows the representation of the evolutionary relations between the proteins, more proteins need to be included in this studies to allow the identification of proteins as TTV1 and also to obtain a different cluster of Cas proteins associated with the CRISPR system.

3.2 *In vivo* Acquisition Assays

3.2.1 CRISPR-Cas system type I-E of *Pseudomonas*

To study acquisition in the type I-E of *Pseudomonas*, the necessary components of the CRISPR array were transformed in *E. coli* BL21-AI cells (see Materials and Methods - *In vivo* acquisition assays). This way, two different plasmids were transformed: one of them encoding the genes codifying for the Cas1-Cas2 complex and another one, the leader and repeat fragments of the CRISPR array (see Figure 5). It was necessary to transform the *cas1* and *cas2* genes since it is known that the BL21-AI cells since it is known that these cells have CRISPR array however, no Cas proteins are encoded in their genome [10]. These cells were then used in this assay since the entire machinery of the CRISPR-Cas system is not present and it allows the detection of, for example, acquisition of spacers from the host genome.

In the PCRs performed to detect acquisition the objective is to amplify the fragment of the CRISPR

array between the leader and the first spacer. In the cases the system shows acquisition, the new spacer acquired and a new repeat (comprising a fragment with around 60 bp in total) would be inserted in this region. This way, in the case the system acquires new spacers, two different populations will be obtained, one in which the CRISPR array have acquired a new spacer and another one in which no new spacers were incorporated in the native CRISPR array. This way, by performing the PCR to detect acquisition, two different bands, corresponding to the amplification of these two different populations, are obtained. In type I-E of *Pseudomonas* the expected size of the non expanded population is approximately 160bp and the size of the expanded population is around 220 bp (Figure 11).

In this study, the PCRs to detect acquisition were performed using degenerate primers binding to the first repeat and an external reverse primer binding to the backbone (in the case of the type I-E, since the *leader* and *repeat* sequences were inserted in a p13SS plasmid, the external reverse primer binds precisely to the p13SS backbone). Degenerate primers were used since the CRISPR array is in a plasmid and consequently, the population of non expanded is too large when compared with the expanded population. This way, a more sensitive method as the use of degenerate primers has to be used in order to detect the presence of the expanded population.

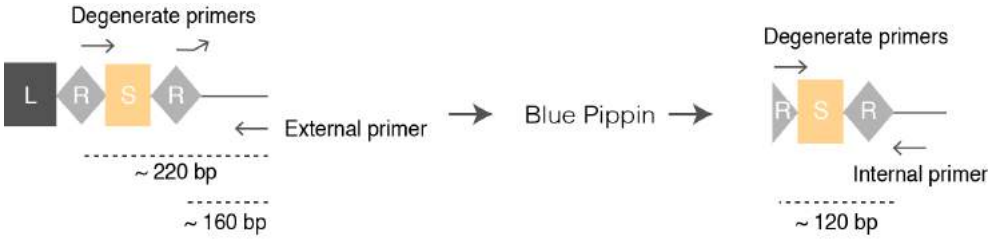

Figure 10: Binding sites of the primers used in the *in vivo* acquisition assays and expected sizes of the expanded and non-expanded bands. In the right PCR, the degenerate forward primers bind to the *repeat* fragment (R) and the external reverse primer binds to the backbone. It is expected to obtain fragments with approximately 160bp in case the non expanded population is amplified and with approximately 220bp in the case the expanded population is amplified. Are also represented the *leader* (L) and *spacer* (S) fragments. In the PCR after the automated gel extraction (Blue Pippin) are used, again, the degenerate forward primers binding to the *repeat* fragment and, as reverse primer, an internal primer binding immediately after the upstream *repeat* sequence. It is expected to obtain with only one size (approximately 120bp, corresponding to the amplification of the expanded CRISPR array.

By analysis of the amplicons obtained in the acquisition PCR it was possible to see, first of all, that no acquisition was detected in the negative control. As expected, no expanded band can be detected since in this PCR was amplified the CRISPR array of BL21-AI cells in which the system was not induced and, consequently, the pCas12 was not overexpressed and no Cas1-Cas2 complex was available to allow the incorporation of spacers in the CRISPR array. However, in all the remaining samples, in the presence or absence of vCas4, it was possible to detect a clear expanded band (+1 band) meaning that in all cases, new spacers were acquired by the CRISPR-Cas system. This allow us to conclude that the *Pseudomonas* CRISPR components transformed in the *E. coli* cells were functional and that, as expected, it is possible to detect acquisition of new spacers.

Regarding the relative intensities of the expanded and non expanded bands, it was possible to see that in all cases, it looks like the amplification of the non expanded population was less intense in the presence of vCas4 than in the empty 2AT control plasmid (Figure11), being this result less evident in the case of KPP25

in which intensities of the expanded band are more or less the same as the intensities verified in the positive control. It means that, in fact, as expected, the vCas4 proteins have shown an effect in the amount of new spacers acquired by the CRISPR-Cas systems: in the presence of this protein, less spacers were acquired.

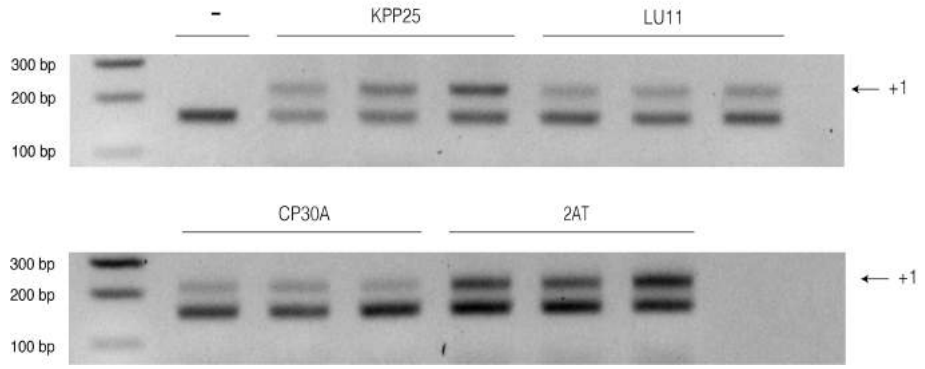

Figure 11: PCR of the *in vivo* acquisition assays in the type I-E of *Pseudomonas* CRISPR-Cas system. Top (right to left): Negative control, KPP25 vCas4 and LU11 vCas4; Down (right to left): CP30A and 2AT empty plasmid (positive control). The band corresponding to the amplification of the expanded (+1) population in which CRISPR array the new spacers were incorporated is marked with a black arrow.

To allow the quantification of this results, the intensities of the non expanded and expanded bands were measured either in the case the vCas4 proteins were or not present (these results can be found in Appendix C. With the values obtained was further quantified the normalized percentage of expanded band in the presence of KPP25, LU11 and CP30A vCas4 proteins and in the absence of this protein (2AT empty plasmid).

By analysis of the obtained percentages of acquisition, it was possible to conclude that no relevant differences on the amount of spacers acquired in the presence of vCas4 KPP25 can be detected. However, in the cases that LU11 and CP30A vCas4 are present, the amount of spacers acquired by the CRISPR-Cas systems decreases significantly being this effect way more evident in the case CP30A vCas4 is present (Figure 12).

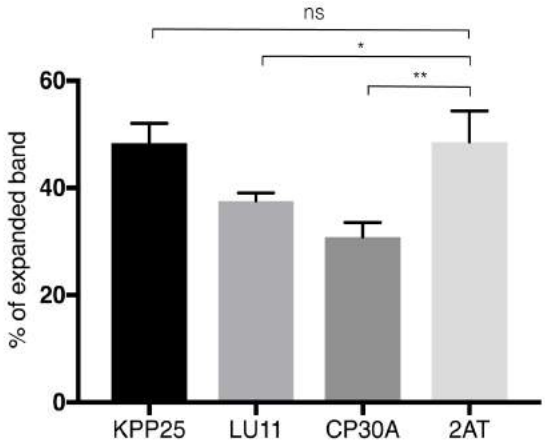

Figure 12: Percentage of the expanded band in the PCR of the *in vivo* acquisition assays in the type I-E of *Pseudomonas* CRISPR-Cas system. Percentages obtained by division of the intensities of the expanded band divided by the sum of the intensities of both expanded and non expanded band for the cases KPP25, LU11 and CP30A vCas4 is present or in the absence of vCas4 (2AT). It is also represented the statistical relevance analysis between the % of expanded band in each one of the cases vCas4 is present in comparison to the case in which this protein is absent (ns: not significant; *:significant; **:highly significant).

With the evidence that the vCas4 proteins in study have an effect in the amount of spacers acquired by the CRISPR-Cas system it was found interesting to understand if the origin of these new acquired spacers would be different in the cases the vCas4 proteins were present.

This way, after the PCR to detect acquisition, was performed an automated gel extraction (Blue Pippin) in order to collect only the population in which new spacers were acquired by selection of the DNA with the size corresponding to the expanded band. The collected DNA was then amplified in an additional PCR. In this PCR, after Blue Pippin, were used the degenerate primers but the reverse primer used was an internal primer binding to the first spacer, making sure that the obtained amplicons contained only new spacers acquired by the CRISPR-Cas system. In the PCR after automated gel extraction, only one band with approximately 120 bp was expected to be seen (Figure11).

It is possible to verify that, as expected, in this PCR only one band corresponding to the amplification of the expanded array is detected with the predicted size of approximately 120 bp. This way, the expanded band was correctly purified and amplified and these fragments can be further sequenced.

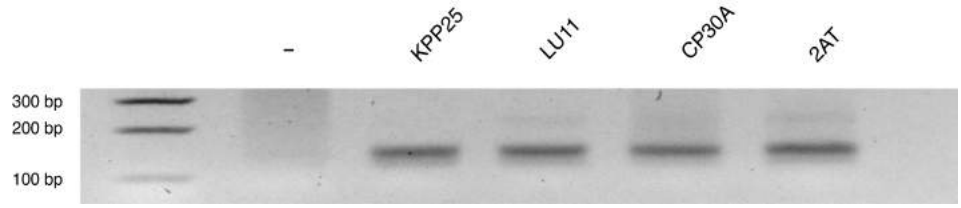

Figure 13: PCR after automated gel extraction of the *in vivo* acquisition assays in the type I-E of *Pseudomonas* CRISPR-Cas system. From right to left: negative control, KPP25 vCas4, LU11 vCas4, CP30A vCas4 and 2AT (positive control).

As explained before, the CP30A viral Cas4 protein was previously studied and it is known to have an influence in the acquisition of novel spacers in the type II-C CRISPR-Cas systems found in *Campylobacter jejuni* bacterial strains. However, curiously, from the three vCAs4 proteins in study, the CP30A was the one showing higher influence in the amount of spacers acquired, when compared with the p2AT empty plasmid. This result is highly interesting since this protein was tested in a CRISPR-Cas system completely different from the host one, meaning that the influence of the vCas4 proteins might not be dependent of the interaction between this protein and the host CRISPR-Cas system but is is universal and possible to be verified in different CRISPR-Cas systems.

This way, in order to understand this non specific interaction between vCas4 and the acquisition in different CRISPR-Cas systems, even if not related with the host, the spacers acquired in the type I-E CRISPR array of *Pseudomonas* were subjected to a preliminary evaluation of their origin.

Thus, the amplicons of the PCR after Blue Pippin of CP30A and the negative control were transformed in pGEM-T vector and transformed in *E. coli* DH5α cells for Blue-White Screening. The white colonies obtained were picked and sent for sequencing. The new spacers acquired were identified and, by BLAST, it was possible to evaluate from where in the plasmids transformed or host genome these spacers came from (Table 10).

In this analysis were evaluated 13 protospacers from KPP25, 12 from LU11, 10 from CP30A and 9 from the positive control. In the cases that KPP25 vCas4, LU11 vCas4 and no vCas4 protein was present (p2AT

Table 10: Percentage of spacers originated from each one of the plasmids p2AT, p13SS and pACYC in the presence of KPP25, LU11 and CP30A vCas4 and in the absence of vCa4 (2AT empty plasmid). Correspondent percentage of spacer with plasmid and genome origin for each one of the conditions in study.

	p2AT	p13SS	pACYC	**% Plasmid**	**%Genome**
KPP25	62	23	15	100	0
LU11	25	50	25	100	0
CP30A	30	0	10	40	60
2AT	33	22	45	100	0

empty plasmid) was verified that none of the novel spacers was originated from the host genome. In these cases, 100% of the novel spacers acquired were originated from plasmids (Table 10). Contrarily to these results, in the case vCas4 CP30A was present, it was verified that, 60% of the novel spacers acquired were originated from the host genome and, the remaining 40%, originated from either the plasmid pTU225 or the plasmid pTU234. This result is a clear evidence that the presence of vCas4 proteins leads to the acquisition of less protospacers by the type I-E CRISPR-Cas system of *Pseudomonas* and that, interestingly, this protospacers are mainly originated from the host genome.

With the evidence that the presence of CP30A vCas4 leads to the incorporation of spacers originated from the genome, was also important to determine the length of the newly acquired spacers and their consensus PAM of the new protospacers acquired (Figure 14).

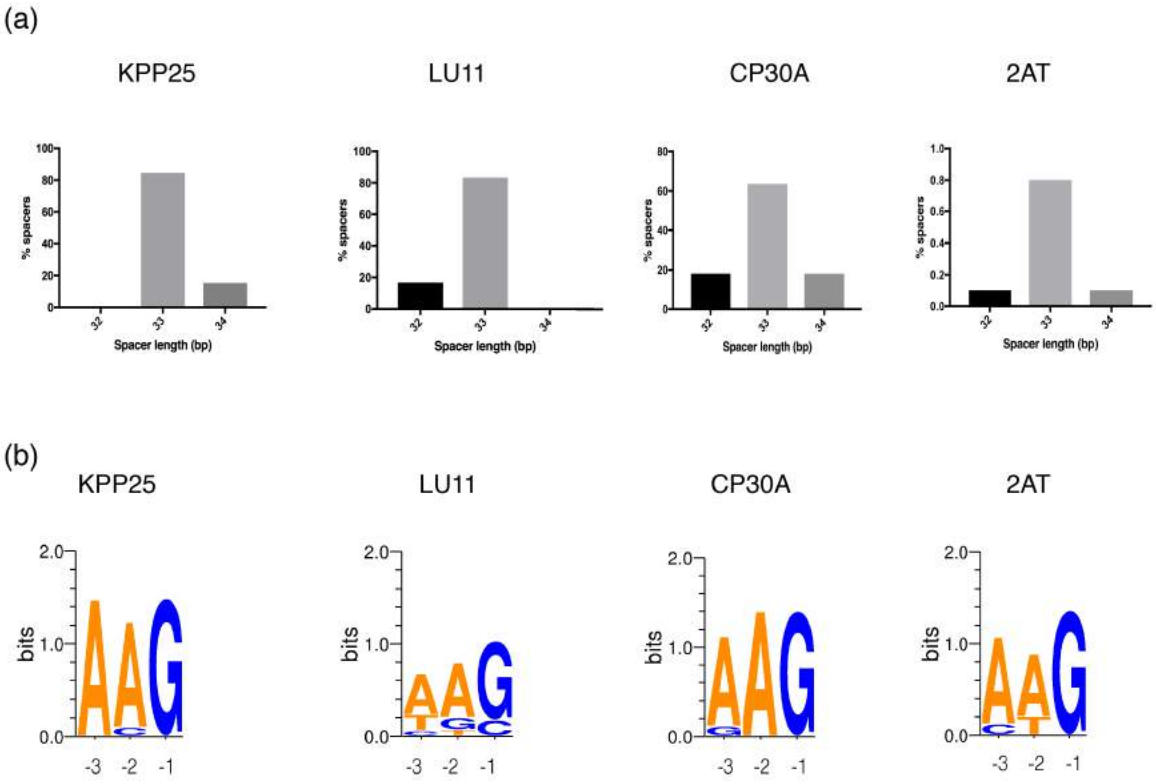

Figure 14: Length and PAM consensus sequence of the novel spacers acquired by the CRISPR array of type I-E of *Pseudomonas* (a) Spacer length distribution in the presence of KPP25, LU11 and CP30A vCas4 and in its absence (2AT empty plasmid). (b) PAM consensus sequence of the novel spacers acquired by the CRISPR array in the presence of KPP25, LU11 and CP30A vCAs4 proteins and in its absence (2AT empty plasmid). Obtained using WebLogo 3.6.0.

It was possible to conclude that the most predominant spacer length of the novel spacers acquired was 33 bp and their consensus PAM is AAG either in the case that vCas4 protein is present or not. Since this spacer length and the AAG PAM are both characteristic of the I-E CRISPR-Cas system of *P. aeruginosa* [49], we can conclude that the vCas4 of this study does not have influence in those parameters.

Since we were testing a type I-E of *Pseudomonas* in BL21-AI cells that already have the CRISPR array of the type I-E of *E.coli* and, moreover, since the repeat sequences of the types I-E of both *E. coli* and *Pseudomonas* are very similar (Figure 15) there was the possibility that the Cas1-Cas2 complex of *Pseudomonas* type I-E couldn't efficiently distinguish between the two repeat sequences, leading to the incorporation of new spacer in the BL21-AI CRISPR array instead of the plasmid containing the type I-E of *Pseudomonas* CRISPR array. This way, as a control of the previously obtained results, was evaluated the acquisition of spacers in the host chromosome and consequently, the possibility of having lost information of novel spacers acquired in the previous assays.

This way, the cells in which was evaluated the acquisition in the type I-E of *Pseudomonas* were subjected to a similar PCR to detect acquisition but using primers binding to the the type I-E of *E. coli* CRISPR array. As template for the negative control were used BL21-AI cells in which this array was expected to be amplified, and consequently, one band corresponded to the non expanded array, obtained. If the Cas1-Cas2 complex was incorporating spacers in the chromosome, it was expected to detect an expanded band. The expected size of the non expanded band was approximately 150 bp and a band with approximately 60 bp more was expected to be seen in the cases spacers were incorporated in the chromosome. It was expected to detect the acquisition of new spacers in this CRISPR-Cas system encoded in type I-E of *Pseudomonas* since a similar experiment was previously performed in the homologous type I-E of *E. coli* by Yosef *et al.* (2012).

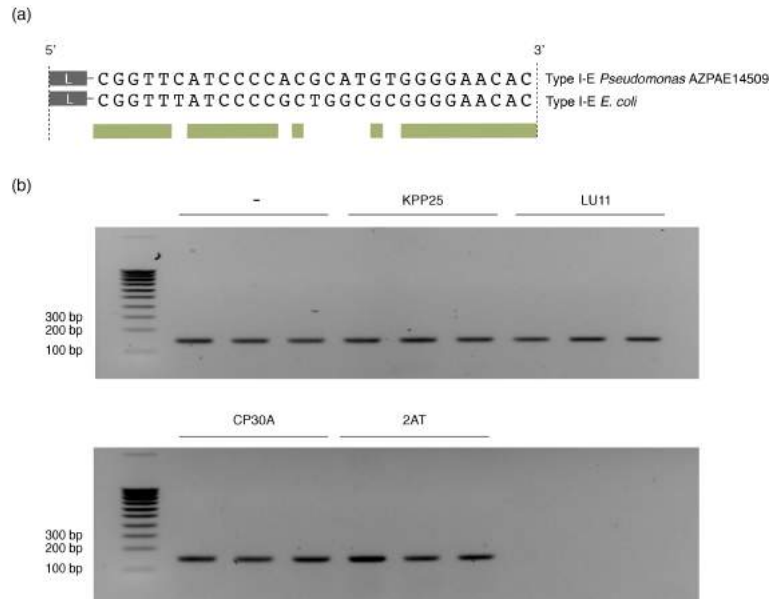

Figure 15: Assay to detect the incorporation of spacers in BL21-AI CRISPR array during *in vitro* activity assays in type I-E CRISPR system of *Pseudomonas*(a) Alignment of repeat sequences of both CRISPR-Cas systems type I-E from *Pseudomonas* and type I-E from *E. coli* (the conserved nucleotides are marked in green). (b) PCR to detect the incorporation of spacers in the *E. coli* BL21-AI CRISPR array in which it is possible to see only one band correspondent to the amplification of the non expanded population by the evaluation of biological triplicates. Top (left to right): negative control, KPP25, LU11; Down (left to right): CP30A and 2AT.

It was possible to conclude that, like in the negative control, only the non expanded band could be detected in all the samples (Figure 15). It means that not a significant number of spacers acquired by the Cas1-Cas2 complex were incorporated in the host CRISPR array and also that no information was lost when evaluating the spacers incorporated in the CRIPSR array of *Pseudomonas*.

3.2.2 CRISPR-Cas system type I-C of *Pseudomonas*

With the results obtained in the *in vivo* acquisition in the type I-E of *Pseudomonas*, the same approach was applied in the type I-C. This study was performed with the objective of understanding if the results obtained in the type I-E were reproducible in different types of CRISPR-Cas systems and if, for example, the effect of CP30A viral Cas4 in acquisition is also verified in the type I-C, supporting the hypothesis that this vCas4 proteins shows universal activity and is not related with the CRISPR-Cas system. Besides that, this study allows us to understand if, even if no significant difference was detected in the amount of spacers acquired by type I-E of *Pseudomonas* in the presence of vCas4 KPP25, this protein has an effect when tested in a different type of host CRISPR-Cas system. Besides that, the evaluation of the type I-C of *Pseudomonas* is very relevant since this system has encoded a Cas4 protein in its CRISPR loci and this way it is also possible to evaluate if this vCas4 supplemented to the system competes with the native one already present.

To study acquisition in the type I-C of *Pseudomonas*, as in the type I-E, the necessary components of the CRISPR array had to be transformed in *E. coli* BL21-AI cells (see Materials and Methods - *In vivo* acquisition assays). The CRISPR-Cas components from type I-C of *Pseudomonas* used in this study (Cas1, Cas2, leader, repeat and Cas4 fragments) were ordered having as template the CRISPR components from the *P. aeruginosa* VA-134 strain. As in type I-E of *Pseudomonas*, were used degenerate primers and an external primer binding to the backbone to detect acquisition.

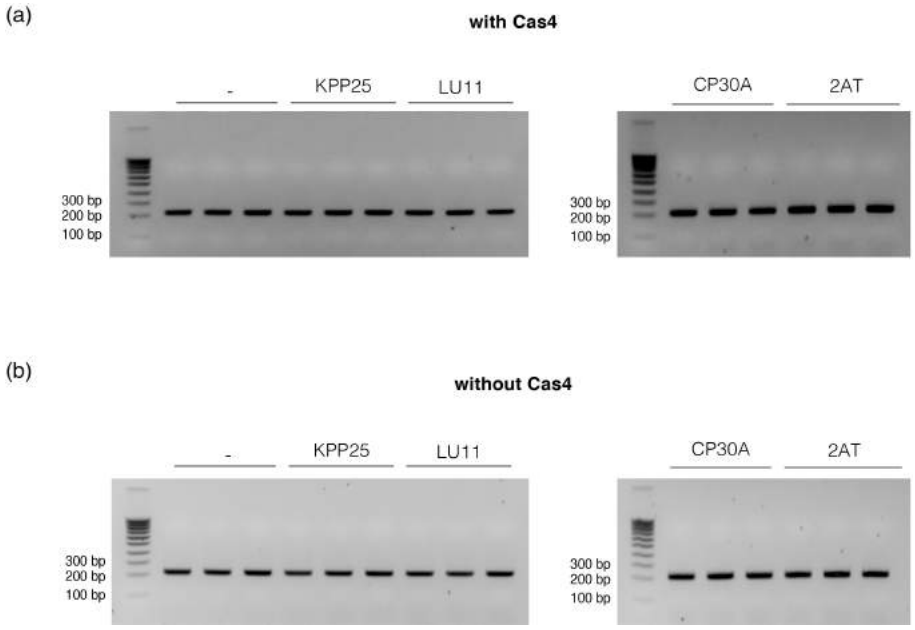

Figure 16: PCR of the *in vivo* acquisition assays in the type I-C of *Pseudomonas* CRISPR-Cas system in the (a) presence and (b) absence of the Cas4 protein from type I-C of *Pseudomonas*. Right to left: Negative control, KPP25 vCas4, LU11 vCas4, CP30A vCas4 and 2AT empty plasmid (positive control). Only the band corresponding to the amplification of the non expanded (+0) population, in which no new spacers were incorporated, is present.

After performing this PCR, it was possible to see that no expanded band is present in any of the samples. It means that the CRISPR components transformed in the BL21-AI cells are not able to incorporate new spacers in the CRISPR array and, since this result was obtained either in the cases the native Cas4 is present or not, we also conclude that the Cas4 is not fundamental to detect acquisition in this CRISPR-Cas type.

In order to confirm this results, the triplicates of each condition were pulled together, an automated gel extraction was performed and the collected DNA was amplified by PCR. This assay was performed with the objective of amplifying the non expanded band and making sure that any acquisition would be detected, even if the amount of novel spacers acquired is very small. In this PCRs was also included a positive control in order to see if the PCR conditions were appropriated since it was expected to not see any band and it was important to make sure that the lack of band was not due to the use of non adequate PCR conditions. The template of the positive control PCR was the plasmid pTU230 (Cas1-Cas2-Leader-Repeat fragment in pACYC).

By analysis of the results it is possible to see that, in fact, no acquisition can be detected since no amplification was obtained in the PCR. By comparing this results with the obtained in the positive control, in which is possible to see a clear band with the expected size of the degenerate primer and first spacer in the CRISPR array (50 bp in total), it is possible to conclude that the negative result obtained is reliable since the PCR conditions might have lead to positive amplification of the CRISPR array (Figure 17).

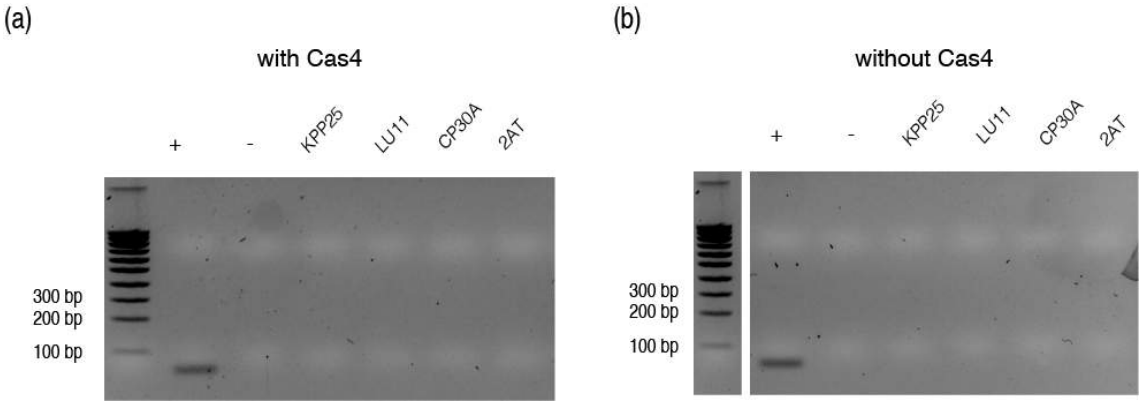

Figure 17: PCR after automated gel extraction of the *in vivo* acquisition assays in the type I-C of *Pseudomonas* CRISPR-Cas system in the (a) presence and (b) absence of the Cas4 protein from type I-C of *Pseudomonas*. Right to left: Negative control, KPP25 vCas4, LU11 vCas4, CP30A vCas4 and 2AT empty plasmid (positive control). No amplification can be detected neither in the presence or absence of vCas4 or in the presence or absence of the Cas4 protein from type I-C of *Pseudomonas* except in the samples corresponding to the positive control corresponding to a the amplification of the CRISPR array in which no automated gel extraction was applied.

Finally, it allows the conclusion that the type I-C of *Pseudomonas* is not able to acquire novel spacers (naive acquisition), either in the case that native Cas4 is present or not, inhibiting the possibility of concluding about vCas4 influence in this CRISPR-Cas system.

Only *in vitro* naive acquisition was previously described in the type I-C CRISPR-Cas systems [28] and no *in vivo* acquisition was ever described. The only *in vivo* acquisition described in this type of CRISPR systems was priming acquisition [49]. Contrarily to naive adaptation, in which spacers that are not already cataloged in the host CRISPR array are there incorporated, the priming acquisition occurs in the case that the CRISPR

system already has memory spacers against an invader [20,35]. It is known that once a host acquires a single spacer against an invader, it becomes more likely to subsequently acquire additional spacers from the area near the priming target region in the phage genome [11]. This process by which pre-existing spacers facilitate rapid spacer acquisition is known as primed spacer acquisition (or priming) [12, 59]. Unlike naive acquisition that only requires the presence of the Cas1-2 complex, the priming acquisition additionally requires the presence of a Cascade (CasA-E), Cas3 and the crRNA [12].

This way, it might happen that only in the presence of the completed CRISPR machinery or in the presence of proteins such as the vCas4 proteins, it would possible to detect acquisition in type I-C of *Pseudomonas*.

3.2.3 Assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-C and I-E of *Pseudomonas*

With the results obtained it was found interesting to understand if the interaction between vCas4 and the CRISPR-Cas systems is direct or a consequence of an indirect influence. This way, since it was verified that the CP30A vCas4 protein has a strong effect in the CRISPR-Cas adaptation and, moreover, that its presence leads to the acquisition of protospacers originated from the genome, further co-purification assays were then performed in order to understand if this protein strongly interacts with the Cas1-Cas2 complex. If it was verified that these proteins co-purifiy and, consequently, that they strongly interact, it allow us to conclude that the vCas4 proteins directly interact with the CRISPR-Cas system and that the incorporation of spacers from the genome might be a result of this protein-protein interaction between the vCas4 proteins and the Cas1-Cas2 complex.

This assay is also interesting since it was already described that the Cas4 proteins encoded in the CRISPR locus have an influence in the incorporation of new spacers by these systems accomplished by the direct interaction of this protein with the Cas1-Cas2 complex [46] and, as found in the phylogenomic studies performed, the vCas4 proteins are highly similar to their homologs from the type I CRISPR-Cas systems. So with this study we can also understand if the known interaction with Cas1-Cas2 complex is lost or not in the vCas4 proteins.

Besides the study of CP30A vCas4 co-purification, this assay was also performed in the case of KPP25 vCas4. This protein was evaluated since it is encoded in a phage that infects proteins knowing to have the types CRISPR-Cas types I-E and I-C and, hereby, this protein shows a higher probability of interacting with the Cas1-Cas2 complex encoded in these types of CRISPR system. Besided that, this protein was also included as a way to better understand the reason why no naive acquisition was detected in the type I-C because, even if no naive acquisition was detected in this type of CRISPR-Cas system even when KPP25 vCas4 was present, it doesn't mean that these protein is not interacting with the system and that it is not necessary to detect acquisition. This analysis allows us to conclude if KPP25 is an important component in the adaptation mechanism of this CRISPR-Cas types or if other proteins are required, instead of this one.

In this assay, *E. coli* BL21-AI cells were transformed with the plasmids carrying the vCas4 proteins in which the His6-SUMO Tag was also attached and also with the plasmid in which the Cas1-Cas2 complex of the types I-E and I-C of *Pseudomonas* were codified. Since the SUMO Tag was attached to the vCas4 proteins it was expected to purify this protein and in the case it strongly interacts with the Cas1-Cas2 complex,

also co-purify at least one of these proteins.

In the SDS-page gel resulting from the co-purification with the type I-E CRISPR components, it was expected to detect a band with around 43kDa corresponding to the expected size of the KPP25 vCas4 protein. In the case of the CP30A vCas4 protein this expected size was 44kDa. In the cases Cas1 and Cas2 were co-purified it was expected to detect bands with 33kDa and 10kDa, respectively. It could also happen that only the His6-SUMO tag is detected and its expected size is around 14kDa.

By analysis of the obtained results it is possible to see that in the case of KPP25 vCas4 protein (Figure 18 a) an intense band between 40 and 55 kDa can be detected. Since the size of this protein is around 43kDA we can conclude that it was positively purified. In the case of CP30A vCas4 protein (Figure 18 b), an intense band can also be seen in the same region of sizes allowing the conclusion that this protein was also positively purified.

In both purifications it was possible to see some additional bands with very low intensity. However, none of these band had the size expected for the Cas1 or Cas2 proteins suggesting that any of these proteins was co-purified along with the vCas4 protein.

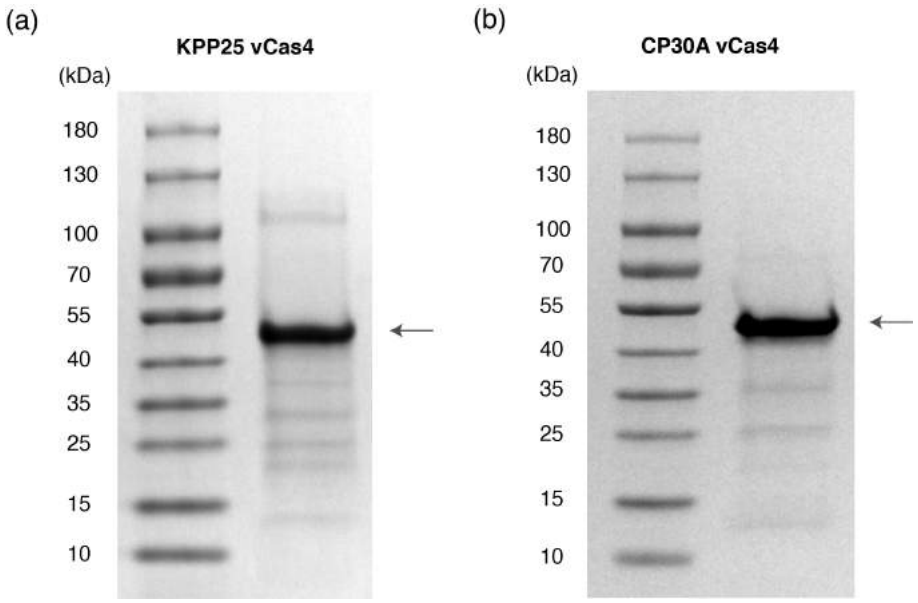

Figure 18: SDS-page gel resulting from the assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-E of *Pseudomonas* in the case of (a) KPP25 vCas4 or (b) CP30A vCas4. The band demonstrating the positive purification of KPP25 vCas4 (43 kDa) and of CP30A vCas4 (44 kDa) is marked with a black arrow. In the cases Cas1 and Cas2 were co-purified it was expected to detect bands with 36 kDa and 11 kDa, respectively.

These samples were additionally evaluated by Mass Spectrometry (see Materials and Methods - Mass Spectrometry) since this method would allow the detection of any possible interaction between the vCas4 protein and the Cas1-Cas2 complex, even if less significant. In this analysis, was evaluated the presence of peptides with the same mass-to-charge ratio of the ones known to be part of each one of the vCas4 proteins and, also, from the Cas1 protein. This way, it was possible to accurately evaluate the presence of each one of these proteins in the samples. In the case of KPP25, 11 peptides were selected and, in the case of CP30A, 12. In both cases, 4 of these peptides were from the His6 SUMO Tag. From Cas1 protein, were also selected 11 peptides.

By analysis of the obtained chromatograms it is possible to conclude that, as expected, both KPP25 and CP30A and their tags are present in the samples. However, for both samples, no clear results of Cas1 also being co-purified can be detected. The peaks detected in the case of Cas1 have very low intensity (1000 times less intensity when compared to the intensities obtained for the peaks in the vCas4 chromatogram) and are located in the noise area (Figure 20).

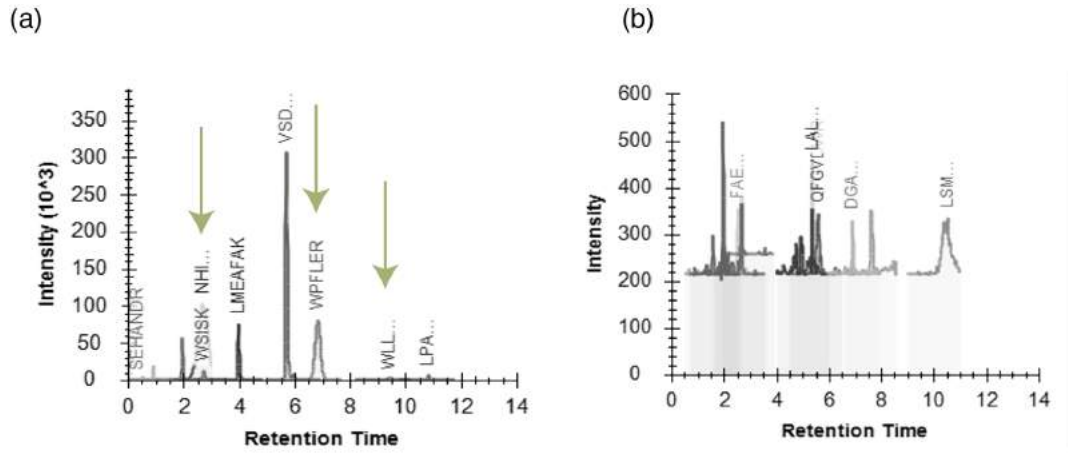

Figure 19: Chromatograms obtained for Mass Spectrometry analysis of Cas1-Cas2 co-purification with KPP25 vCas4 (a) Chromatogram obtained for the detection of peptides from KPP25 vCas4 protein (b) Chromatogram obtained for the detection of peptides from the Cas1 protein. In green are marked the positive peptides detected in each one of the samples that do not correspond to the His6 SUMO Tag.

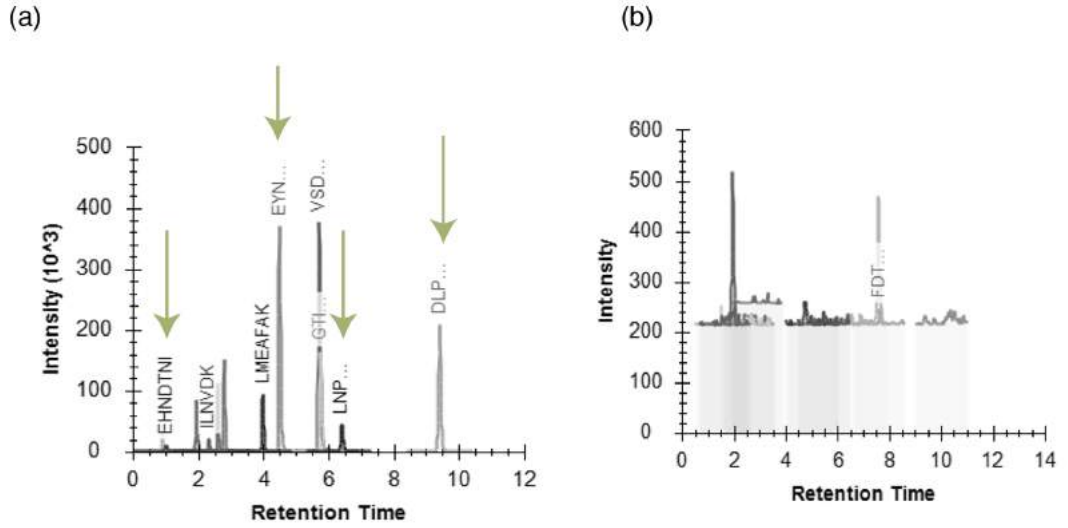

Figure 20: Chromatograms obtained for Mass Spectrometry analysis of Cas1-Cas2 co-purification with CP30A vCas4 (a) Chromatogram obtained for the detection of peptides from CP30A vCas4 protein (b) Chromatogram obtained for the detection of peptides from the Cas1 protein. In green are marked the positive peptides detected in each one of the samples that do not correspond to the His6 SUMO Tag.

In the co-purification of the type I-C, if the Cas1 and Cas2 proteins co-purified it was expected to detect a band in the molecular weight of 36kDa and 11 kDa.

In the purification of KPP25 vCas4 protein (Figure 21) it was possible to detect an intense band in between 40 and 55 kDa which means that as in type I-E, this protein was positively purified. No additional bands can

be detected in the SDS-page gel meaning that no significant proteins co-purifying can be detected. The same conclusion can be taken from the purification of CP30A vCas4. Since no co-purification was detected in the type I-E of *Pseudomonas* and the results obtained in the SDS-page gels presented didn't show promising co-purification in the type I-C, these samples were not additionally evaluated by Mass Spectrometry.

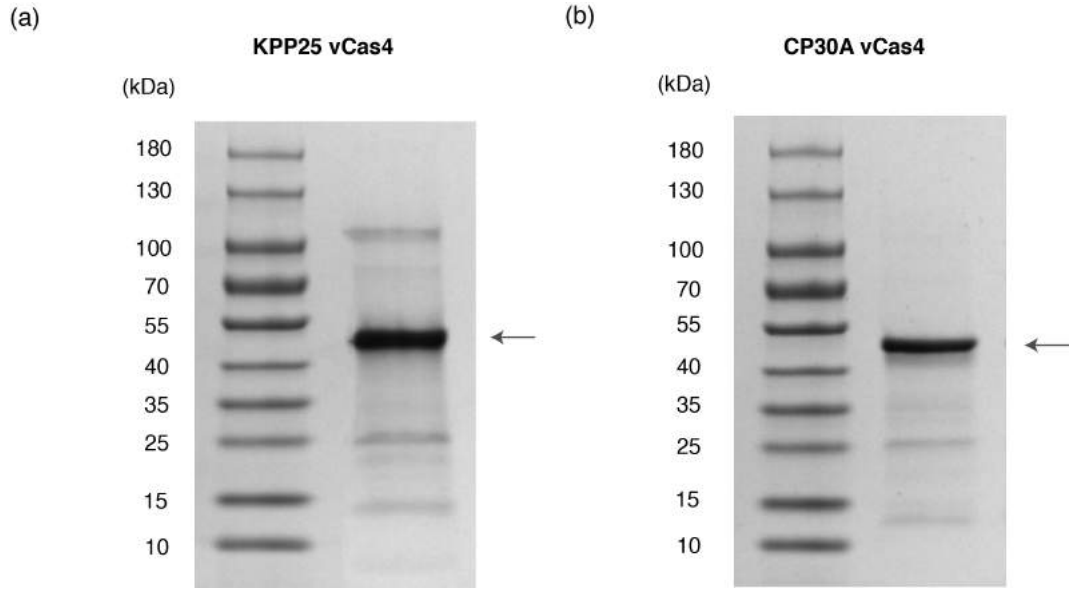

Figure 21: SDS-page gel resulting from the assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-C of *Pseudomonas* in the case of (a) KPP25 vCas4 or (b) CP30A vCas4. The band demonstrating the positive purification of KPP25 vCas4, with approximately 43 kDa is marked with a black arrow in the left gel and the band demonstrating the positive purification of CP30A vCas4, with approximately 44 kDa is marked with a black arrow in the right gel. In the cases Cas1 and Cas2 were co-purified it was expected to detect bands with 36kDa and 11kDa which is not verified.

Taken together, these results allows us to conclude that the vCas4 proteins do not interact with the Cas1-Cas2 complex.

3.3 Biochemistry Assays

Since it was possible to conclude that the vCas4 does not interact directly with the CRISPR-Cas system, further biochemistry assays were performed in order to understand the possible vCas4 protein activity and, finally, understand possible indirect interaction between this protein and the CRISPR-Cas system mechanism.

These way, the all the vCas4 proteins were first purified using Ni-NTA Affinity Chromatography (see Materials and Methods - Protein Purification).

By analysis of the SDS-page gels obtained, in the case of KPP25 vCas4 it was possible to see a clear band with size between 40 and 55 kDa and the expected size of this protein is 43kDa. In the case of LU11 vCas4 protein which expected size is 57kDa, a band of this size was observed. Finally, in the case of CP30A vCas4 protein, a band with size between 40 and 55 kDa can be detected and the expected size of this protein is around 44 kDa. Since in all cases, bands were obtained with the expected sizes of the given proteins, we could conclude that they all were successfully purified by Ni-NTA affinity chromatography (Figure 22).

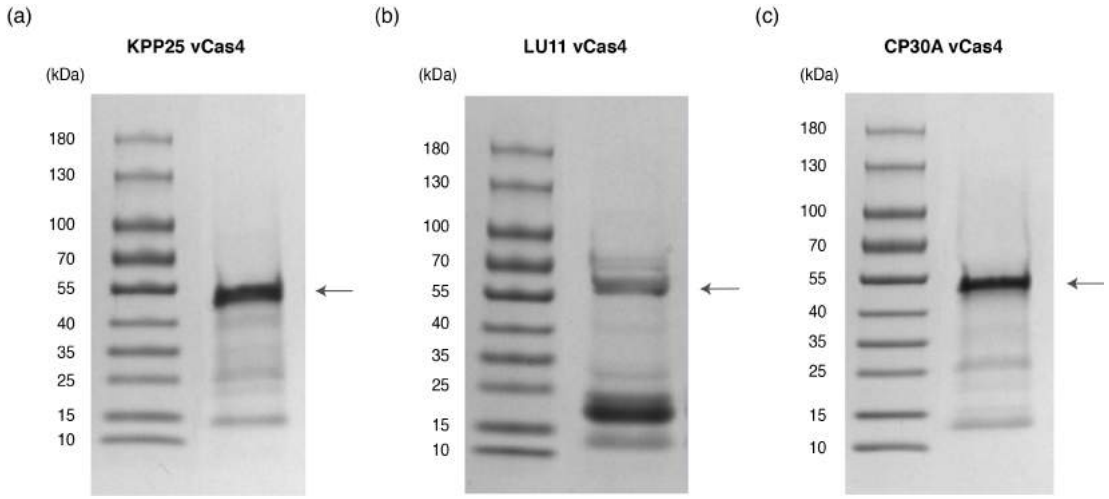

Figure 22: SDS-page gel obtained after Ni-NTA protein purification of (a) KPP25 vCas4 (b) LU11 vCas4 and (c) CP30A vCas4. Marked with a black arrow are the expected sizes of each one of the vCas4 proteins. Since a clear band is possible to be detected with the same size of the marked expected sizes, it is possible to conclude that the vCas4 proteins were purified adequately.

The purified vCas4 proteins obtained from Ni-NTA affinity chromatography were further subjected to an additional size-exclusion chromatography (see Materials and Methods - Protein Purification). This way, any additional protein contamination present in the obtained elutions were eliminated making sure that any further *in vitro* assays were performed only in the presence of the vCas4 proteins in study.

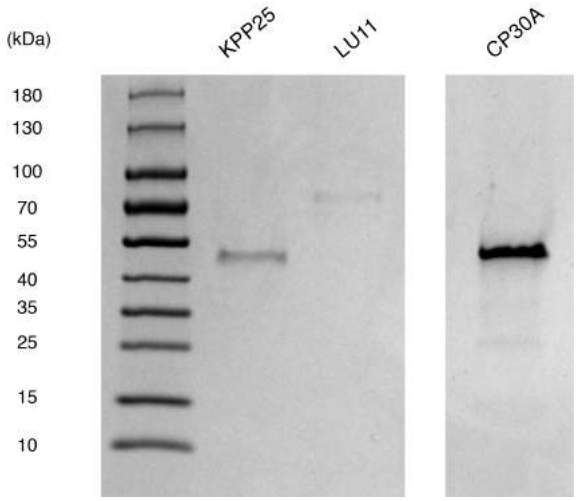

Figure 23: SDS-page gel after size exclusion. SDS-page that allows visualization of purified KPP25 and CP30A vCas4 proteins after size exclusion.

After size exclusion it is possible see clear bands with approximately 45kDa in the gels obtained for both KPP25 and CP30A vCas4 proteins (Figure 23). Since the expected sizes of these proteins was, respectively, 43 and 44kDa we can conclude these vCas5 proteins were correctly purified by this methodology. It is also possible to see that in the case of LU11 vCas4 protein, no clear band with the expected size of this protein can be detected meaning that the protein purification of this vCAs4 was not successfully performed. This way, the further *in vitro* assays were performed only in the evaluation of KPP25 and CP30A vCas4 protein activity.

Moreover, it is possible to see that for both samples, only one clear band is present. This was, as wanted, we can conclude that the proteins of interest were purified from all remaining proteins and they are completely isolated. Consequently, further *in vitro* acquisition assays could be performed knowing that the protein activity detected would be caused only by the presence of the vCas4 proteins in study and not other contamination proteins present in elution. This way, were then performed the assays to evaluate the *in vitro* activity of KPP25 and CP30A vCas4.

In the first assay performed to detect nuclease activity of the vCas4 proteins, these proteins were incubated with different types of DNA (linear double-stranded DNA, circular double-stranded DNA and circular single-stranded DNA) during two hours at 37ºC in the presence of two different buffers ($MgCl_2$ and $MnCL_2$). As negative control, the DNA samples were incubated in the presence of buffer but without the addition of any protein. In this case, since no protein was present, no degradation of DNA was expected to be seen. In another hand, in the samples in which the vCas4 proteins are present, if these proteins have nuclease activity, it is expected to detect the degradation of DNA that can be seen in the visualization of the incubated DNA in an agarose gel being that in the case the DNA is degraded, a smear band is present in the gel.

With the results obtained was possible to see that the presence of vCas4 proteins didn't lead to the degradation of dsDNA, neither linear or circular (Figure 24). The dsDNA samples used were pACYC plasmid, either linearized or in its natural circular conformation. In the first three samples of each gel it is possible to see a clear band with around 4kbp corresponding to the known size of this plasmid. Since this clear band is seen also in the presence of vCas4 proteins it is possible to conclude that no degradation occured. In the three samples of each gel in which circular double-stranded DNA was loaded it was possible to see a band with lower size that in the linear dsDNA samples. Even if the same plasmid was used in both cases, they do not show the same migration pattern in the agarose gel since it is known that circular DNA easily migrates when separated by electrophoresis [44]. This way, as expected, a band with lower molecular weight is seen in the case of circular dsDNA. Once again, since a clear band is seen also in the presence of vCas4 proteins it is possible to conclude that no degradation of circular dsDNA occurred.

The template used to detect nuclease activity in single-stranded circular DNA was M13DNA. The size of this bacteriophage DNA is known to be around 6kbp. In the case of ssDNA samples it is possible to see that in the presence of KPP25 vCas4 proteins a small smear can be detected in both buffers being more evident in the case of $MnCl_2$ buffer. This suggests that KPP25 might have nuclease activity and this activity is enhanced by the presence of the $MnCl_2$ buffer.

In the case that CP30A vCas4 was incubated with circular ssDNA, the presence of DNA can't be detected in the agarose gel (Figure 24). This result suggests that this protein is actively degradating the DNA sample and that this strong nuclease activity might lead to the degradation of the DNA sample into single nucleotides that can't be stained using the DNA loading dye. The bromophenol blue and xylene cyanol stains present in the DNA Loading dye might not be able to bind to such small fragments of DNA explaining the loose of DNA sample detected in both gels. In order to confirm this hypothesis, the same *in vitro* activity assay were repeated over time and with additional supplementation of EDTA to completely stop the reactions. EDTA (ethylenediaminetetraacetic acid) is a chelating agent that binds to metal complexes inhibiting their activity that was used in this assay with the objective of stopping the degradation reactions of CP30A vCas4 proteins

over the DNA template [45].

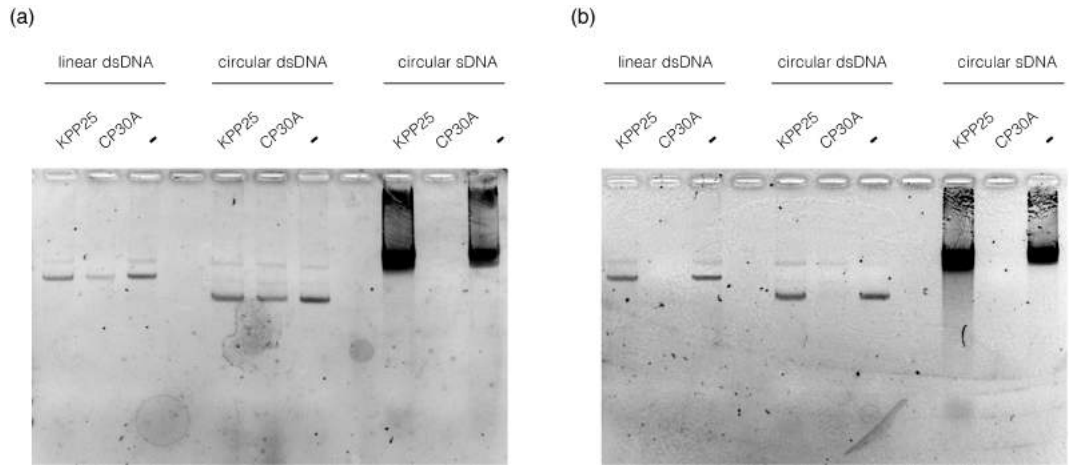

Figure 24: *In vitro* assay for determination of vCas4 protein activity. Both KPP25 and CP30A vCas4 proteins were incubated with different types of DNA (linear double-stranded DNA, circular double-stranded DNA and circular single-stranded DNA) during two hours at 37ºC in the presence of (a) $MgCl_2$ buffer and (b) $MnCl_2$ buffer.

Since the preliminary results obtained for CP30A haven't shown degradation of double-stranded DNA by this protein, not even after two hours of incubation, the first assay over time was preformed incubating the proteins of interest only with single-stranded circular DNA (M13 DNA). The reactions were stopped right after incubation and after 5, 10, 30 and 60 minutes of incubation with the presence of either $MnCl_2$ and $MgCl_2$ buffer (Figure 24).

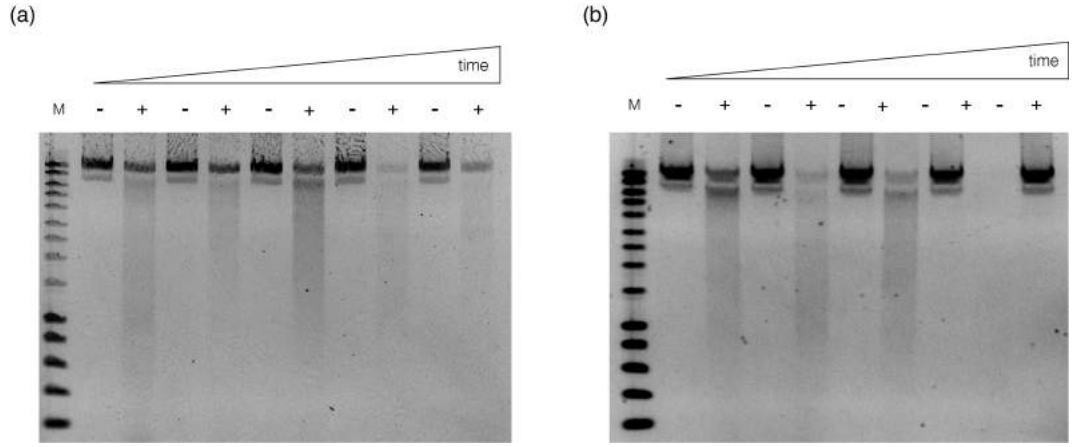

Figure 25: *In vitro* assay for determination of CP30A vCas4 protein activity over time. CP30A vCas4 proteins were incubated with circular single-stranded DNA over time (0, 5, 10, 30 and 60 minutes) at 37ºC in the presence of (a) $MgCl_2$ buffer and (b) $MnCl_2$ buffer. After incubation, the reactions were stopped by EDTA supplementation.

This study allow us to see that the nuclease activity of CP30A vCas4 protein can be detected even when the reaction is immediately stopped after incubation of this protein with the DNA sample. The smear resulting from the nuclease activity of DNA degradation is seen after 5 and 10 minutes in both buffers. However, after 30 and 60 minutes (and specially in the case $MnCl_2$ buffer is supplemented) the detection of DNA in the agarose gel is lost. This might be caused by the reason previously explained: since the CP30A vCas4 protein shows high activity, single nucleotides are generated inhibiting the DNA loading dye binding and

consequent visualization. Once again, higher activity was demonstrated in the presence of $MnCl_2$ buffer instead of $MgCl_2$.

To make sure that this protein has no nuclease activity over double-stranded DNA, the same conditions were then testes and the *in vitro* activity assays were performed again but with addition of EDTA immediately after incubation and after 10 and 30 of reaction, for all the DNA samples tested before.

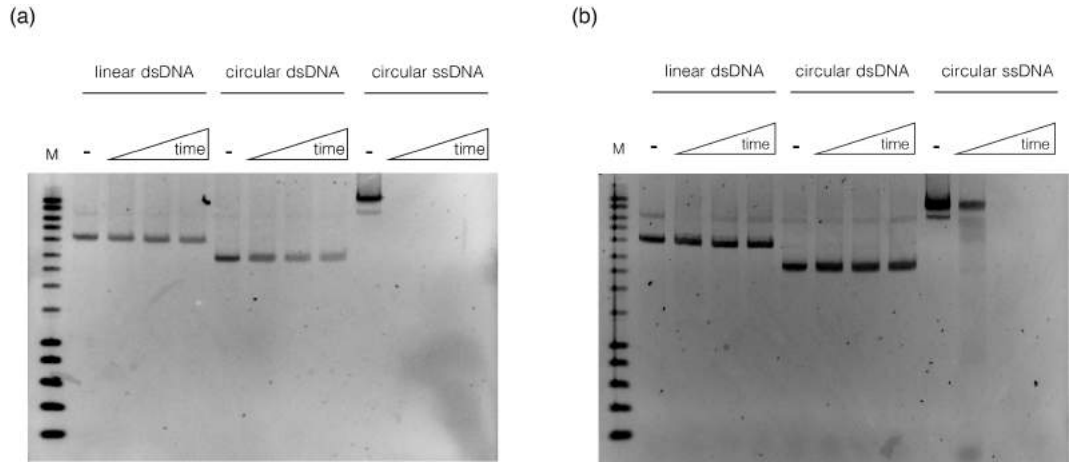

Figure 26: *In vitro* assay for determination of CP30A vCas4 protein activity over time. CP30A vCas4 proteins were incubated with with different types of DNA (linear double-stranded DNA, circular double-stranded DNA and circular single-stranded DNA) over time (0, 10 and 30minutes) at 37ºC in the presence of (a) $MgCl_2$ buffer and (b) $MnCl_2$ buffer. After incubation, the reactions were stopped by EDTA supplementation.

The obtained results allow the conclusion that, in fact, CP30A vCas4 is a nuclease that, as KPP25 vCas4 has preference for single-stranded DNA. No activity and degradation of DNA was demonstrated neither in the presence of linear or circular double-stranded DNA. When comparing both vCas4 proteins in study it is possible to conclude that CP30A shows enhanced activity when compared to the KPP25 since this activity can be detected with lower incubation time and only if the reaction is stopped by supplementation of EDTA. Moreover, it is also possible to conclude that in both proteins in study the proteins are more active in the presence of $MnCl_2$ instead of $MgCl_2$ buffer. It means that the nuclease activity of this vCas4 protein is metal dependent and that this activity is coordinated by a mechanism that preferentially involves manganese ion (Mn_{2+}) coordination at the active site (instead of magnesium ion coordination). Previous experiments were done demonstrating that the Cas4 family nucleases were metal dependent [72] which proofs again the similarities between the already known Cas4 proteins associated to the CRISPR-Cas system and the Cas4 homologs encoded in phage genomes, under the scope of this study.

With clear evidence that vCas4 proteins have endonuclease activity, since it cleaves circular ssDNA, further assays to determine if this protein is an exonuclease were performed in the case of CP30A vCas4 protein using both $MnCl_2$ and $MgCl_2$ buffers and two different DNA substrates, one incorporating a fluorescent label at the 3'-end and another incorporating the fluorescent label at the 5'-end.

First of all, it was possible to see that, as expected, no degradation of the labelled oligonucleotides occur in the absence of vCas4 protein (Figure 27). Besides that it is possible to see that CP30A vCas4 cleaves the 5'-end labelled oligonucleotide in the presence of both Mg_{2+} and Mn_{2+} into three products with different sizes corresponding to the three more intense bands obtained in the gel (more evident in the case MnCl2 was

used as buffer and as expected since it is already known that the presence of this metal enhances protein activity). In another hand, it was verified that cP30A vCas4 cleaved the 3'-end labelled into a product with a single size without any observable intermediates.

The partial degradation of the 5'-end labelled oligonucleotides by the vCas4 protein in three fragments suggest that this protein is degrading DNA in specific position. This evidence and the fact that it was previously demonstrated that this protein is a nuclease with preference for circular single-stranded DNA allow us to conclude that CP30A vCas4 is an endonuclease.

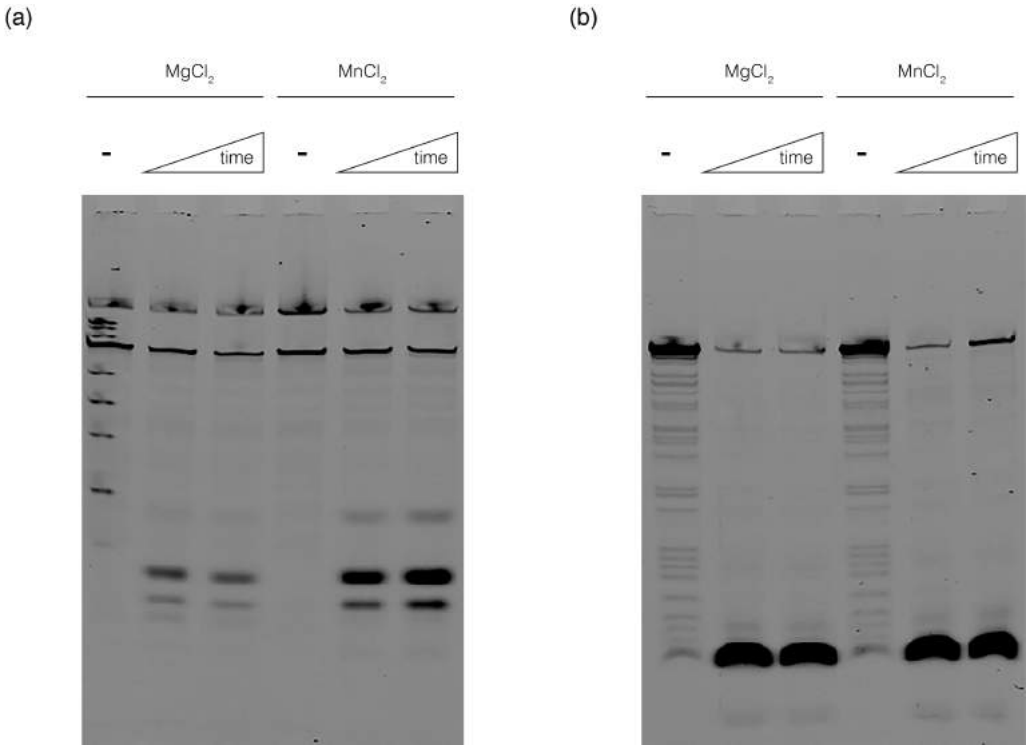

Figure 27: 20% SDS-page gel to detect exonuclease activity of CP30A vCas4 over (a) 5'-end labelled oligonucleotides or (b) 3'-end labelled oligonucleotides. Both reactions were performed in the presence of both $MnCl_2$ and $MgCl_2$ buffers and over time (10 and 30 minutes). After incubation, the reactions were stopped by EDTA supplementation.

Chapter 4

Discussion

The Cas4 proteins has long been implicated in immune adaptation, forming complex with Cas1 [28] and selecting and orienting PAM-compatible spacers. [23,55]. All these recent studies have increased the interest in revealing the biological role of Cas4 in CRISPR-Cas systems. However, these proteins can also be found not associated with the CRISPR-cas loci, being present *solo* in MGEs such as in bacteriophages [18].

Two studies were already done in order to understand the role of these vCas4 proteins and, interestingly, two different completely roles were demonstrated for this protein. In one hand, in the *Thermoporteus tenax virus* the *cas4* gene was split and codifies a coat protein [25]. In other hand, it was shown that CP30A vCas4 is responsible for stimulating the acquisition of host-derived spacers in type II-C CRISPR-Cas systems [16, 17]. These completely disparate roles have then motivated the study of these vCas4 proteins since a lot of questions remain unanswered: first, what are the similarities and differences between these Cas4 proteins encoded in phages and the ones associated with the CRISPR-Cas systems?; Second, does these vCas4 proteins also have a role in CRISPR adaptation as the one encoded in the CRISPR locus?; Third, what is the activity exhibited by these vCas4 proteins and what is the role of this activity and this protein in the phages in which they are encoded?; In this study, were addressed all these questions and shown that, as the Cas4 proteins associated with the CRISPR-Cas systems, vCas4 also has an influence in CRISPR adaptation. We selected three vCas4: KPP25, CP30A and LU11. vCas4 KPP25 is encoded in a phage that infects *Pseudomonas aeruginosa*, specie that encode three CRISPR-Cas systems: I-E, I-C and I-F. We choose the DNA adaptation sequences from type I-E of *P. aeruginosa* AZPAE14509, strain in which CRISPR array were found spacers against the *Pseudomonas* phage KPP25 [7]. This means that this strain is probably naturally infected by the KPP25 phage, allowing us to expect a possible interaction between its vCas4 and its CRISPR adaptation module. Since non spacers were described against the phages here studied in the I-C system, we performed the assays using the sequences from *P. aeruginosa* VA-134 strain. No assays were done in the type I-F CRISPR-Cas systems since only priming acquisition was detected in this system type [1, 50]. In the case of CP30A it was already studied. Its influence in the acquisition of new spacer in type II CRISPR-Cas systems was known being, however, unknown if its influence is specific or not. This way, we expected to understand if the same result can also be detected in CRISPR-Cas types I-E and I-C of *P. aeruginosa* that are, not only different in their constitution, but also not the CRISPR-Cas system of the natural host. Contrarily to KPP25 and CP30A vCas4 and since no CRISPR-Cas system is known in the natural host of LU11 phage (*P. putida*), it is interesting and harder to predict which would be the expected influence of this protein in the

CRISPR-Cas adaptation.

Despite the expected results, the *in vivo* acquisition assays performed in the type I-E CRISPR-Cas system revealed a non significant acquisition in the presence of KPP25 vCas4 protein. However, in assays using CP30A and LU11 viral Cas4, a decrease in the amount of spacers was detected. This results is very interesting since none of the proteins in which the influence was significant is encoded in phages that infect strain with type I-E CRISPR-Cas systems. Moreover, in the case of CP30A, as was already demonstrated in type II-C CRISPR-Cas system, we detect by sequencing a stimulation of acquisition of host-derived spacers in the host CRISPR array. Both results suggest that this phenomenon is not host-CRISPR related and probably, the influence of the CP30A vCas4 is universal: promotes acquisition of host genome derived spacers in any CRISPR-Cas system leading to possible autoimmunity events.

Then, in order to understand how the vCas4 enhances the autoimmunity, biochemistry assays were performed. Since all Cas4 proteins have a conserved RecB nuclease domain, the fact that the vCas4 used in this study are nucleases can be the explanation for the previous results obtained regarding its influence in CRISPR adaptation. Assays performed using circular ssDNA and linear ssDNA demonstrated that KPP25 and CP30A vCas4, presents a ssDNA endonuclease activity, enhanced in the case of CP30A. And this degradation of DNA might be the key explanation for the enhancement of host derived spacer acquisition: since, as more host genome DNA is present in the cells, would be produced more genome fragments than can be used by Cas1-Cas2 because is not described that CRISPR-Cas system has a mechanism to differentiate their own DNA from foreign one. However, this can also be the reason why less spacers were incorporated in the CRISPR array when this protein was present, since DNA fragments created by CP30A vCas4 are not optimal to be integrated by Cas1-2 and in consequence, less acquisition rate was detected. This hypothesis is in concordance with the results obtained with LU11 and with the activity of Cas4 in CRISPR systems. Recently, it was described that Cas4 nuclease activity participates in cleavage of 3' overhangs of protospacers [51] and this processing ensures the formation of optimal protospacers [28]. So, it can be that this activity is maintained in CP30A and LU11 vCas4 proteins and they also make overhangs in the phage derived protospacers inhibiting their recognition by the Cas1-Cas2 complex. The nuclease activity demonstrated by the vCas4 proteins might also be the explanation why no decrease in the amount of spacers acquired was demonstrated in the presence of KPP25 vCas4. Since it was proved by the *in vitro* activity assays that the nuclease activity of KPP25 is more limited than in the case of CP30A, it might have reduced the possibility of producing or modification of protospacers.

In this study, we also confirm that the influence of vCas4 is not related with CRISPR protein interactions. Our results demonstrate that our vCas4 do not interact with the I-E and I-C acquisition module, since non co-purification of Cas1-Cas2 was detected. Consequently, the effect observed and described here, along with the universality of the vCas4 activity, suggests that, is not CRISPR specific related. Even if this protein was initially acquired by phages as, probably, a way to repair DNA during their life-cycle, the collateral activity described in this study, ended up giving them an advantage over bacteria.

4.1 Future Applications

Besides of the gained insights on the vCas4 phylogenomics, interference in CRISPR adaptation and protein activity, this study can have some practical future applications. The enhanced efficiency of phage infection resulting from the presence of vCas4 proteins has as main biotechnological application the improvement of phage therapy effectiveness.

Phage Therapy

The emergence of pathogenic bacteria resistant to most, if not all, of the antimicrobial agents available has become a critical problem in modern medicine. The concern that humankind is reentering the "preantibiotics" era has become real, and the development of alternative antibacterial methods has become one of the highest priorities of modern medicine and biotechnology [**?**, 41, 58].

Prior to the discovery and widespread use of antibiotics, it was suggested that bacterial infections could be prevented and/or treated by the administration of bacteriophages [58].

One of the main advantages of the use of bacteriophages instead of antibiotics to treat bacterial infections is that bacteriophages are very specific to their hosts. This can minimize the chance of secondary infections when compared to the use of antibiotics since the last ones do target both pathogens and normal flora of patients, which can cause secondary infections and eventually superinfections. Besides that, another advantage of the use of bacteriophages is that they replicate at the site of infection contrarily to antibiotics that travel throughout the body. Lastly, bacteriophages are environmentally friendly and are based on natural selection, isolating and identifying bacteria in a very rapid process compared to new antibiotic development, which may take several years and may also not be very cost effective [**?**, 36].

Even if phage therapy might look like a promising alternative to the use of antibiotics, these two therapies have one thing in common. Although the dynamics may differ, the evolution of bacterial resistance to a particular phage, just as to an antibiotic, is inevitable. As described before, the resistance to infection by a phage may involve several different mechanisms in which it's also included the resistance mediated by the CRISPR-Cas systems [31, 35]. Notably, however, contrary to antibiotics, phages will themselves be under evolutionary selection to overcome the new resistance. New phage types evolve continuously and resistance to the bacterial immune systems might result from that [35].

Some ways to tackle the resistance problem have been under the scope of investigation for years. This study have reveled to have a role on that improvement of phage resistance. The discovery of phages in which viral Cas proteins were encoded and, moreover, the founding that these proteins have a clear effect in enhancing host CRISPR autoimmunity can not only be applied as a way to overcome CRISPR-Cas immunity but also in a way to use phages to treat bacterial infections. This way, phages encoding these vCas4 proteins can be used as a better and more efficient tool in phage therapy. Besides that, since it was proofed that vCas4 enhances host CRISPR autoimmunity and, more important, that the observed effect is not CRISPR related but conserved in different types of CRISPR-Cas systems, it means that these vCas4 proteins can also be engineered in different phages in order to increase host autoimmunity.

In the case that LU11 vCas4 was present we also observed a decrease in the amount of novel spacers acquired by the CRISPR-Cas system of type I-E of *Pseudomonas* however, no differences were detected

in the origin, length or PAM of these new spacers acquired. Since the presence of this protein leads to the incorporation of correct invader derived spacer this protein might be further used as a regulator of acquisition in cases it is necessary to reduce the amount of new spacers acquired by the CRISPR-Cas systems.

4.2 Future Work

Even if this study allowed us to gain insight in vCas4 activity, some questions still remain to be answered: First, if vCas4 activity is not related with the CRISPR-Cas system, what is the role of this protein in the phage replication and why does phages evolved to acquire these Cas4 homologs in their genome? Second, since vCas4 leads to the incorporation of host derived spacers in the CRISPR array, what is the further consequences of this phenomena in the later stages of the the CRISPR-Cas mechanism and how do they do that? Regarding the results obtained in this study, we can propose a model that might answer this last question.

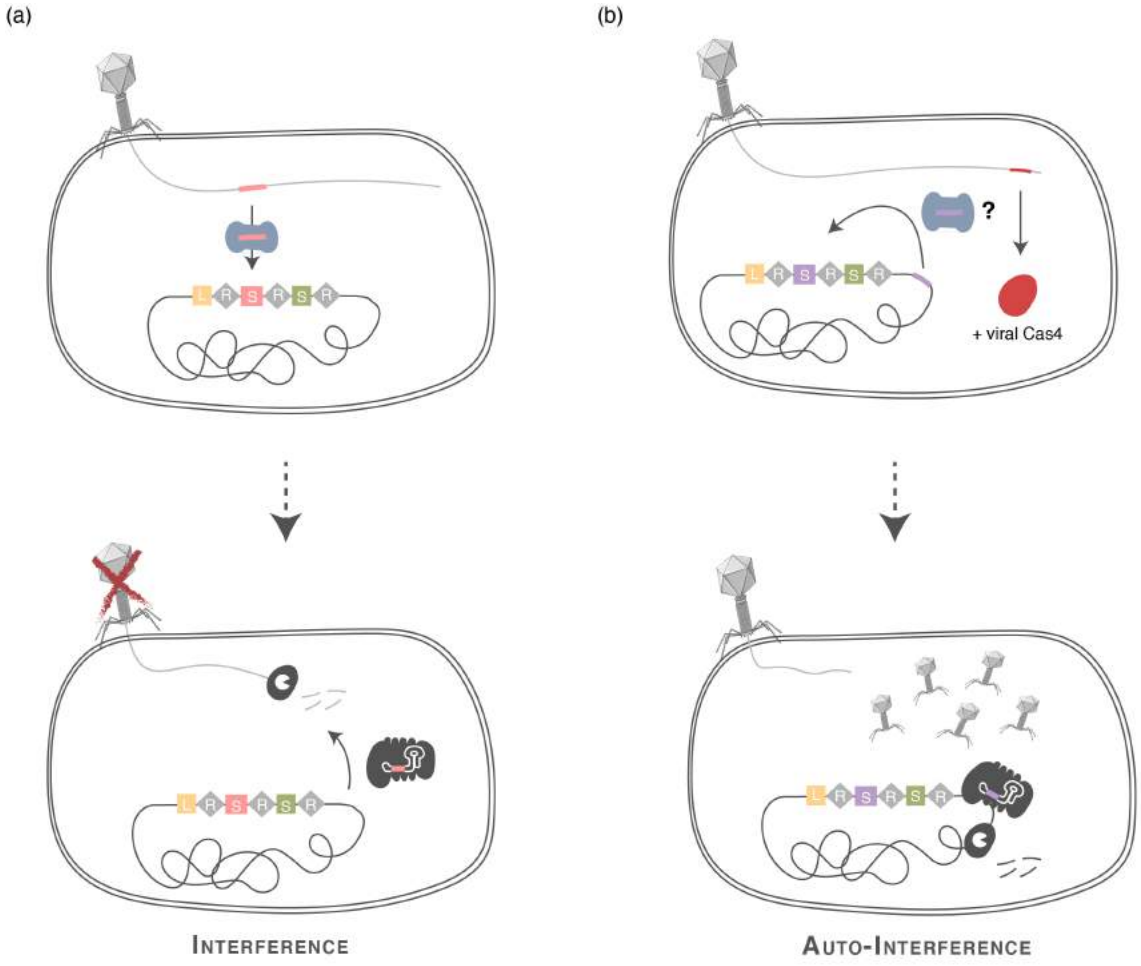

Figure 28: Model explaining vCas4 interference with CRISPR system. (a) CRISPR system mechanism in the absence of of vCas4. The interference mechanisms are activated leading to the destruction of the invader by the incorporation of a protospacer acquired from its genome. (b) CRISPR system mechanism in the presence of vCas4. The Cas1-Cas2 complex incorporates host derived spacers. Since these protospacers have the correct PAM, the presence of the vCas4 proteins will result in auto-interference and lead to positive phage replication and survival. In the CRISPR array are represented the *leader* (L), *repeat* (R) and *spacer* (S) fragments.

In the already known mechanism of interference, in the case the vCas4 proteins are not present, it is known that the interference mechanisms are activated leading to the destruction of the invader by the incorporation of a protospacer acquired from the invader genome in the CRISPR array and consequent expression and processing of its components.

So, we hypothesize that in the case the vCas4 is present, as demonstrated in this study, the Cas1-Cas2 complex incorporates spacers in the CRISPR array which origin is the bacterial genome. As a consequence, the expressed and processed crRNA-effector complexes will recognize and consequently activate the interference mechanisms leading to its destruction. In this case, the presence of the vCas4 proteins will result in auto-interference and lead to positive phage replication and survival.

This study can also motivate the investigation of other Cas protein homologs encoded in phage genomes in a similar fashion and, also, the study of vCas4 influence in different CRISPR-Cas types. As in this study revealing vCas4 properties, the discovery of similar proteins would enhance the knowledge in CRISPR-Cas resistance and, consequently, in the ways to overcome the wide-spread bacterial infections. From the protein alignment of the vCas4 proteins obtained in this study we could, for example, evaluate the effect of the Cas4 homologs encoded in other *Campylobacter* phages as, for example, the vCas4 protein encoded in the *Campylobacter* phage NCTC12673 which protein sequence is highly similar to the CP30A vCas4. Besides that, due to their high diversity, some VCas4 proteins encoded in *Mycobacterium* phages could also be analyzed. Besides that, it could also be interesting to study an homolog that, contrarily to the ones suggested, does have both the lysine and aspartic acid conserved, however, not all the four cysteines are conserved. That study would allow the determination of each residues are responsible for the observed protein activity of vCas4 proteins in the case they lead to an enhancement of host CRISPR autoimmunity.

Besides that, the studies of both KPP25 and LU11 vCas proteins still need to be continued being that in these cases, the origin of the novel spacers acquired still has to be determined. Further biochemistry studies also have to be performed in the presence of vCas4 LU11.

Chapter 5

Conclusion

Here were studied the Cas4 protein homologs encoded in phage genomes according to three different approaches: bioinformatic analysis, *in vitro* acquisition assays and biochemistry studies.

From the bioinformatic analysis it was possible to obtain a database of 112 Cas4 homolog proteins encoded in phage genomes. From the alignment of these vCas4 proteins was possible to conclude they are highly similar to Cas4 proteins associated with type I CRISPR-Cas systems since the four cysteines and additional lysine and aspartic acid residues known to be highly conserved in the family of Cas4 nucleases were also shown to be conserved in the vCas4 proteins. This analysis have also shown that the protein alignment complemented to the localization of the genes encoding for vCas4 proteins in the phage genome and phage life cycle is a better tool to spot proteins that are highly similar to Cas4 but evolved to have different function, when compared to sequence similiarity dendograms. After that, and using the previous alignment, three vCas4 proteins encoded in the Pseudomonas KPP25, Pseudomonas LU11 and Campylobacter CP30A phages were chosen to perform further experimental assays.

In one hand, *in vivo* acquisition assays were performed using I-E system in presence or absent of vCas4. Was observed that the presence of LU11 and CP30A vCas4 proteins decreases the amount of spacers incorporated in the CRISPR-Cas system and in the case of CP30A vCas4 leads to the incorporation of host derived spacers. However, for all the proteins in study, it was demonstrated that the spacers acquired in their presence have the correct length and PAM. On the other hand, in the studied performed in the CRISPR-Cas type I-C of *Pseudomonas*, no naive acquisition was detected neither in the presence or absence of the native Cas4 protein. This result was also not influenced by the presence of vCas4 proteins.

In order to understand the mechanisms underlying the results observed in acquisition, were additionally performed assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-C and I-E of *Pseudomonas*. It was possible to conclude that the vCas4 proteins in study do not interact with the Cas1-Cas2 complexes encoded in these systems. It allowed the conclusion that the verified influence of vCas4 in CRISPR adaptation is not due to a direct interference of this proteins to the components being part of the system.

Since it was demonstrated that the effect of vCas4 proteins in CRISPR-Cas acquisition was not motivated by a direct interaction between the vCas4 proteins and the Cas1-Cas2 complex, further biochemistry assays were performed in order to understand it. This assays allowed the purification of all the vCas4 proteins in study using Ni-NTA affinity chromatography being that, additionally, KPP25 and CP30A were further

subjected to an additional purification step by size-exclusion and were further subjected to *in vitro* assays to determine their activity. The *in vitro* activity assays allow us to conclude that the vCas4 proteins in study have nuclease activity with preference for ssDNA. No activity and degradation of DNA was demonstrated neither in the presence of linear or circular dsDNA. Additionally, it was also possible to conclude that the nuclease activity of this vCas4 proteins is metal dependent and coordinated by a mechanism that preferentially involves manganese (Mn2+) instead of magnesium ions (Mg_{2+}). In addition to that, further assays were performed using CP30A vCas4 in order to verify if the protein is an exonuclease. With the results obtained in all the biochemistry assays it was possible to conclude that, in fact, CP30A vCas4 is a ssDNA endonuclease.

Taken together these results suggest that the nuclease activity of vCas4 might be the reason behind the enhancement of host CRISPR autoimmunity and also that this effect is universal and not CRISPR-Cas specific related. Thus, besides of the gained insights on the vCas4 proteins and its role in CRISPR adaptation, this study opens a door in the possibilities of improving phage therapy effectiveness or use engineered phages a tool to tackle the emergence of pathogenic bacteria resistance problem.

Bibliography

[1] Cristóbal Almendros and Francisco J.M. Mojica. Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. *Nature Microbiology*, 1(8), 2016.

[2] 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(5819):1709–12, 2007.

[3] David Bikard and Luciano A. Marraffini. Innate and adaptive immunity in bacteria: Mechanisms of programmed genetic variation to fight bacteriophages, 2012.

[4] New England Biolabs. *5-alpha E. coli*, (accessed August 6, 2018). `https://international.neb.com/products/c2987-neb-5-alpha-competent-e-coli-high-efficiency`.

[5] 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(1):60–70, 2015.

[6] 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. *Cancer Epidemiology Biomarkers and Prevention*, 321(5891):960–4, 1993.

[7] Kyle C. Cady, Joe Bondy-Denomy, Gary E. Heussler, Alan R. Davidson, and George A. O'Toole. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages, 2012.

[8] 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 and Development*, 22(24):3489–96, 2008.

[9] F. Desiere, R. D. Pridmore, and H. Brussow. Comparative genomics of the late gene cluster from Lactobacillus phages. *Virology*, 275(2):294–305, 2000.

[10] C. Díez-Villaseñor, C. Almendros, J. García-Martínez, and F. J.M. Mojica. Diversity of CRISPR loci in Escherichia coli. *Microbiology*, 156(5):1351–61, 2010.

[11] P. C. Fineran, M. J. H. Gerritzen, M. Suarez-Diez, T. Kunne, J. Boekhorst, S. A. F. T. van Hijum, R. H. J. Staals, and S. J. J. Brouns. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences*, 111(16):E1629–E1638, 2014.

[12] Peter C. Fineran and Emmanuelle Charpentier. Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information, 2012.

[13] Geneious. *Fast and accurate multiple sequence alignment with MAFFT*, (accessed August 22, 2018). `https://www.geneious.com/plugins/mafft-plugin/`.

[14] Bei Gong, Minsang Shin, Jiali Sun, Che-Hun Jung, Edward L. Bolt, John van der Oost, and Jeong-Sun Kim. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proceedings of the National Academy of Sciences*, 111(46):16359–64, 2014.

[15] Rachel E. Haurwitz, Martin Jinek, Blake Wiedenheft, Kaihong Zhou, and Jennifer A. Doudna. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, 329(5997):1355–8, 2010.

[16] Steven P.T. Hooton, Kelly J. Brathwaite, and Ian F. Connerton. The bacteriophage carrier state of Campylobacter jejuni features changes in host non-coding RNAs and the acquisition of new host-derived CRISPR spacer sequences, 2016.

[17] Steven P.T. 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*, 6(JAN):1–9, 2015.

[18] Sanjarbek Hudaiberdiev, Sergey Shmakov, Yuri I. Wolf, Michael P. Terns, Kira S. Makarova, and Eugene V. Koonin. Phylogenomics of Cas4 family nucleases. *BMC Evolutionary Biology*, 17(1):1–14, 2017.

[19] 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(12):5429–33, 1987.

[20] 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(6333), 2017.

[21] 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(6):1565–75, 2002.

[22] 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 and Molecular Biology*, 18:529–536, 2011.

[23] Sebastian N. Kieper, Cristóbal Almendros, Juliane Behler, Rebecca E. McKenzie, Franklin L. Nobrega, Anna C. Haagsma, Jochem N.A. Vink, Wolfgang R. Hess, and Stan J.J. Brouns. Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Reports*, 22(13):3377–3384, 2018.

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

[25] Mart Krupovic, Virginija Cvirkaite-Krupovic, David Prangishvili, and Eugene V. Koonin. Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. *Biology Direct*, 10(1):2–7, 2015.

[26] Tim Künne, Sebastian N. Kieper, Jasper W. Bannenberg, Anne I.M. Vogel, Willem R. Miellet, Misha Klein, Martin Depken, Maria Suarez-Diez, and Stan J.J. Brouns. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Molecular Cell*, 63(5):852–64, 2016.

[27] Simon J. Labrie, Julie E. Samson, and Sylvain Moineau. Bacteriophage resistance mechanisms, 2010.

[28] Hayun Lee, Yi Zhou, David W. Taylor, and Dipali G. Sashital. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Molecular Cell*, 5(1):48–59, 2018.

[29] Ryan T. Leenay, Kenneth R. Maksimchuk, Rebecca A. Slotkowski, Roma N. Agrawal, Ahmed A. Gomaa, Alexandra E. Briner, Rodolphe Barrangou, and Chase L. Beisel. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell*, 62(1):137–47, 2016.

[30] Sofia Lemak, Boguslaw Nocek, Natalia Beloglazova, Tatiana Skarina, Robert Flick, Greg Brown, Andrzej Joachimiak, Alexei Savchenko, and Alexander F. Yakunin. The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: Crystal structure and nuclease activity. *Nucleic Acids Research*, 42(17):11144–11155, 2014.

[31] Bruce R. Levin and James J. Bull. Population and evolutionary dynamics of phage therapy. *Nature Reviews Microbiology*, 2(2):166–73, 2004.

[32] 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(7548):505–510, 2015.

[33] 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(11):722–36, 2015.

[34] Luciano A. Marraffini and Erik J. Sontheimer. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea, 2010.

[35] McGinn Jon Marraffini, Luciano A. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nature Reviews in Microbiology*, pages 960–4, 2018.

[36] Shigenobu Matsuzaki, Mohammad Rashel, Jumpei Uchiyama, Shingo Sakurai, Takako Ujihara, Masayuki Kuroda, Masahiko Ikeuchi, Toshikazu Tani, Mikiya Fujieda, Hiroshi Wakiguchi, and Shosuke Imai. Bacteriophage therapy: A revitalized therapy against bacterial infectious diseases. *Journal of Infection and Chemotherapy*, 11(5):211–9, 2005.

[37] F. J M Mojica, Cesar Díez-Villaseñor, Elena Soria, and Guadalupe Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria, 2000.

[38] F. J.M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, 155(Pt 3):733–40, 2009.

[39] 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(2):174–82, 2005.

[40] Rihito Morita, Shuhei Nakane, Atsuhiro Shimada, Masao Inoue, Hitoshi Iino, Taisuke Wakamatsu, Kenji Fukui, Noriko Nakagawa, Ryoji Masui, and Seiki Kuramitsu. Molecular mechanisms of the whole DNA repair system: A comparison of bacterial and eukaryotic systems, 2010.

[41] Franklin L. Nobrega, Marnix Vlot, Patrick A. de Jonge, Lisa L. Dreesens, Hubertus J. E. Beaumont, Rob Lavigne, Bas E. Dutilh, and Stan J. J. Brouns. Targeting mechanisms of tailed bacteriophages. *Nature Reviews Microbiology*, 2018.

[42] James K. Nuñez, Lucas B. Harrington, Philip J. Kranzusch, Alan N. Engelman, and Jennifer A. Doudna. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, 527(7579):535–8, 2015.

[43] James K. Nuñez, Philip J. Kranzusch, Jonas Noeske, Addison V. Wright, Christopher W. Davies, and Jennifer A. Doudna. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Structural and Molecular Biology*, 21(6):528–34, 2014.

[44] Ariella Oppenheim. Separation of closed circular DNA from linear DNA by electrophoresis in two dimensions in agarose gels. *Nucleic Acids Research*, 9(24):6805–12, 1981.

[45] Claudia Oviedo and Jaime Rodríguez. EDTA: The chelating agent under environmental scrutiny, 2003.

[46] 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(10):2491–500, 2012.

[47] Welkin H. Pope. Expanding the diversity of mycobacteriophages: Insights into genome architecture and evolution. *PLoS ONE*, 6(1), 2011.

[48] Promega. *pGEM®-T and pGEM®-T Easy Vector Systems Technical Manual*, (accessed August 22, 2018). https://nld.promega.com/resources/protocols/technical-manuals/0/pgem-t-and-pgem-t-easy-vector-systems-protocol/.

[49] Chitong Rao, Denny Chin, and Alexander W. Ensminger. Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA*, 23(10):1525–1538, 2017.

[50] 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(13):8516–26, 2014.

[51] Clare Rollie, Shirley Graham, Christophe Rouillon, and Malcolm F. White. NAR breakthrough article: Prespacer processing and specific integration in a type I-A CRISPR system. *Nucleic Acids Research*, 46(3):1007–1020, 2018.

[52] Marius Rutkauskas, Tomas Sinkunas, Inga Songailiene, Maria S. 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*, pages S2211–1247, 2015.

[53] Julie E. Samson, Alfonso H. Magadán, Mourad Sabri, and Sylvain Moineau. Revenge of the phages: Defeating bacterial defences, 2013.

[54] Thermo Fischer Scientific. *BL21-AI One Shot Chemically Competent E. coli*, (accessed August 6, 2018). https://www.thermofisher.com/order/catalog/product/C607003.

[55] Masami Shiimori, Sandra C. Garrett, Brenton R. Graveley, and Michael P. Terns. Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Molecular Cell*, 70(5):814–824, 2018.

[56] Kyle C. Smith, Eduardo Castro-Nallar, Joshua N.B. Fisher, Donald P. Breakwell, Julianne H. Grose, and Sandra H. Burnett. Phage cluster relationships identified through single gene analysis. *BMC Genomics*, 14(410), 2013.

[57] Adi Stern and Rotem Sorek. The phage-host arms race: Shaping the evolution of microbes, 2011.

[58] A. Sulakvelidze, Z. Alavidze, and J. G. Morris. Bacteriophage Therapy. *Antimicrobial Agents and Chemotherapy*, 45(3):649–659, 2001.

[59] 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(4), 2012.

[60] Mark D. Szczelkun, Maria S. Tikhomirova, Tomas Sinkunas, Giedrius Gasiunas, Tautvydas Karvelis, Patrizia Pschera, Virginijus Siksnys, and Ralf Seidel. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences*, 111(27):9798–803, 2014.

[61] Stineke van Houte, Angus Buckling, and Edze R. Westra. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiology and Molecular Biology Reviews*, 80(3):745–63, 2016.

[62] 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(4):840–53, 2015.

[63] Edze R. Westra, Ekaterina Semenova, Kirill A. Datsenko, Ryan N. Jackson, Blake Wiedenheft, Konstantin Severinov, and Stan J.J. Brouns. Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genetics*, 9(9), 2013.

[64] Edze R. Westra, Paul B.G. van Erp, Tim Künne, Shi Pey Wong, Raymond H.J. Staals, Christel L.C. Seegers, Sander Bollen, Matthijs M. Jore, Ekaterina Semenova, Konstantin Severinov, Willem M. de Vos, Remus T. Dame, Renko de Vries, Stan J.J. Brouns, and John van der Oost. CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Molecular Cell*, 46(5):595–605, 2012.

[65] 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(7365):486–489, 2011.

[66] Addison V. Wright and Jennifer A. Doudna. Protecting genome integrity during CRISPR immune adaptation. *Nature Structural and Molecular Biology*, 23(10):876–883, 2016.

[67] Addison V. Wright, Jun Jie Liu, Gavin J. Knott, Kevin W. Doxzen, Eva Nogales, and Jennifer A. Doudna. Structures of the CRISPR genome integration complex. *Science*, 357(6356):1113–1118, 2017.

[68] Yibei Xiao, Sherwin Ng, Ki Hyun Nam, and Ailong Ke. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature*, 550(7674):137–141, 2017.

[69] Chaoyou Xue, Arun S. Seetharam, Olga Musharova, Konstantin 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(22):10831–47, 2015.

[70] Ido Yosef, Moran G Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research*, 40(12):5569–76, 2012.

[71] Xinyan Yu, Yue Xu, Yu Gu, Yefei Zhu, and Xiaoqiu Liu. Characterization and genomic study of "phiKMV-Like" phage PAXYB1 infecting Pseudomonas aeruginosa. *Scientific Reports*, 7(1):13068, 2017.

[72] 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(10), 2012.

Appendix A

gBlocks Gene Fragments Sequence

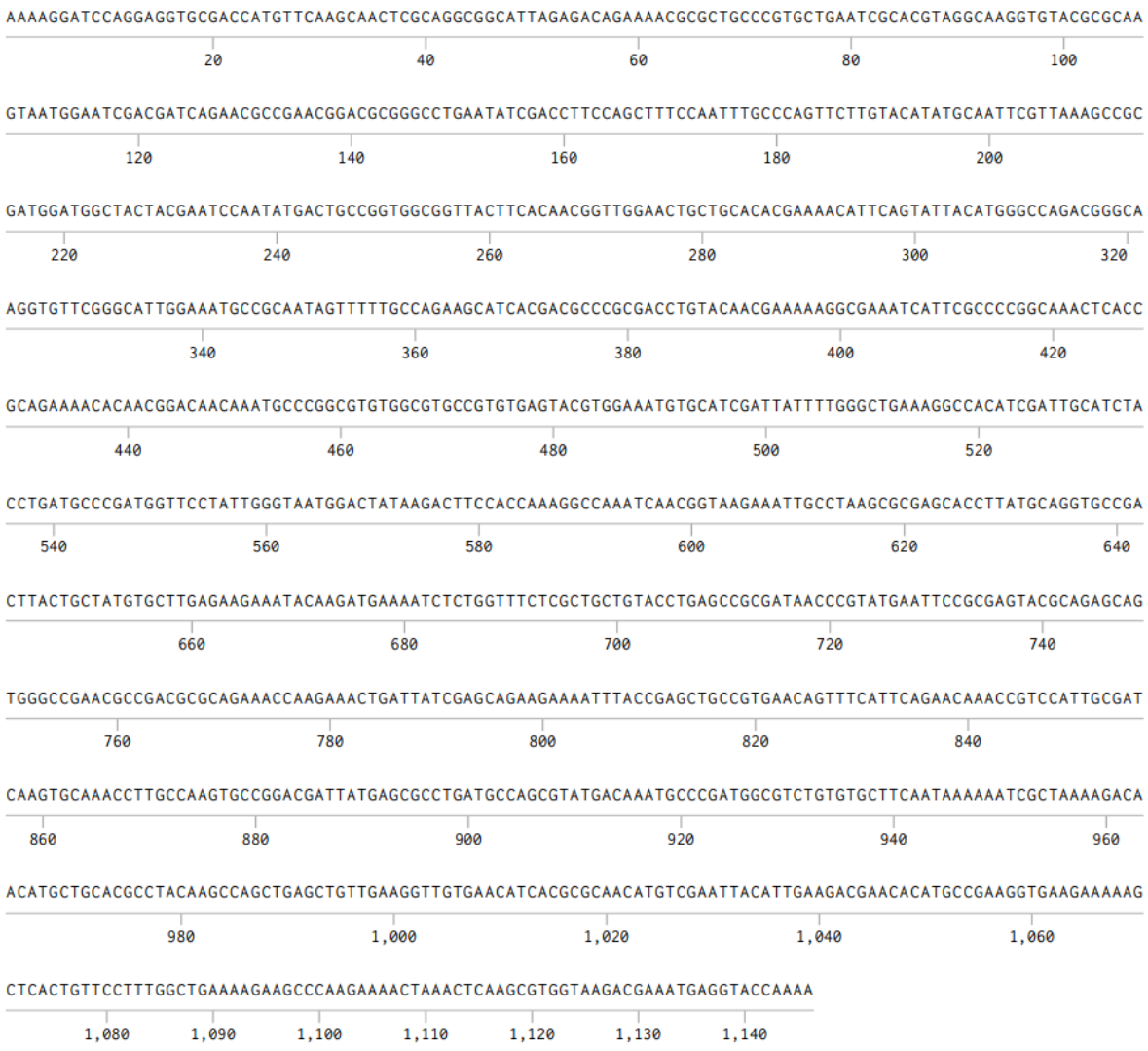

Figure 29: gBlocks sequence of Cas4 homolog gene encoded in LU11.

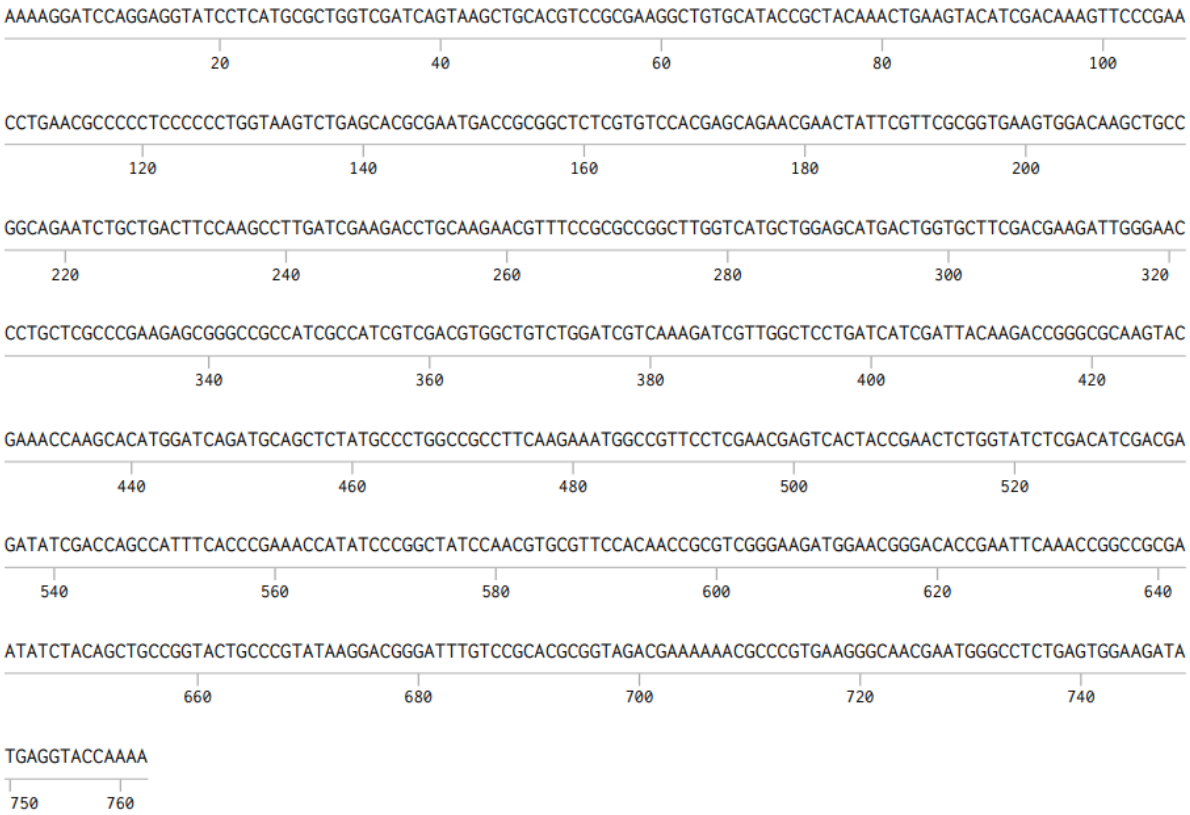

Figure 30: gBlocks sequence of Cas4 homolog gene encoded in KPP25.

Figure 31: gBlocks sequence of Cas4 homolog gene encoded in CP30A.

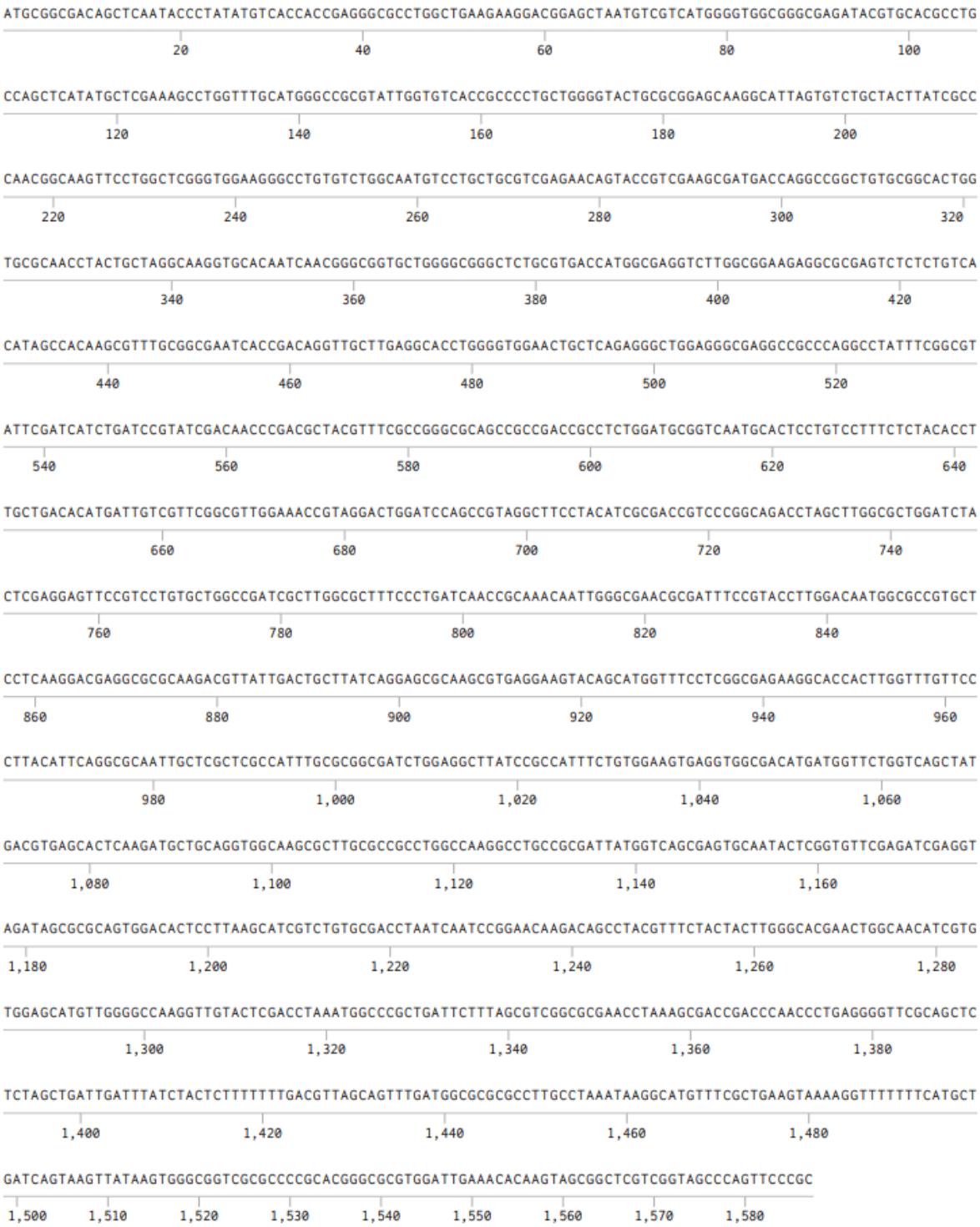

Figure 32: gBlocks sequence of Cas1-Cas2-Leader-Repeat genes from type I-C CRISPR system of *Pseudomonas*.

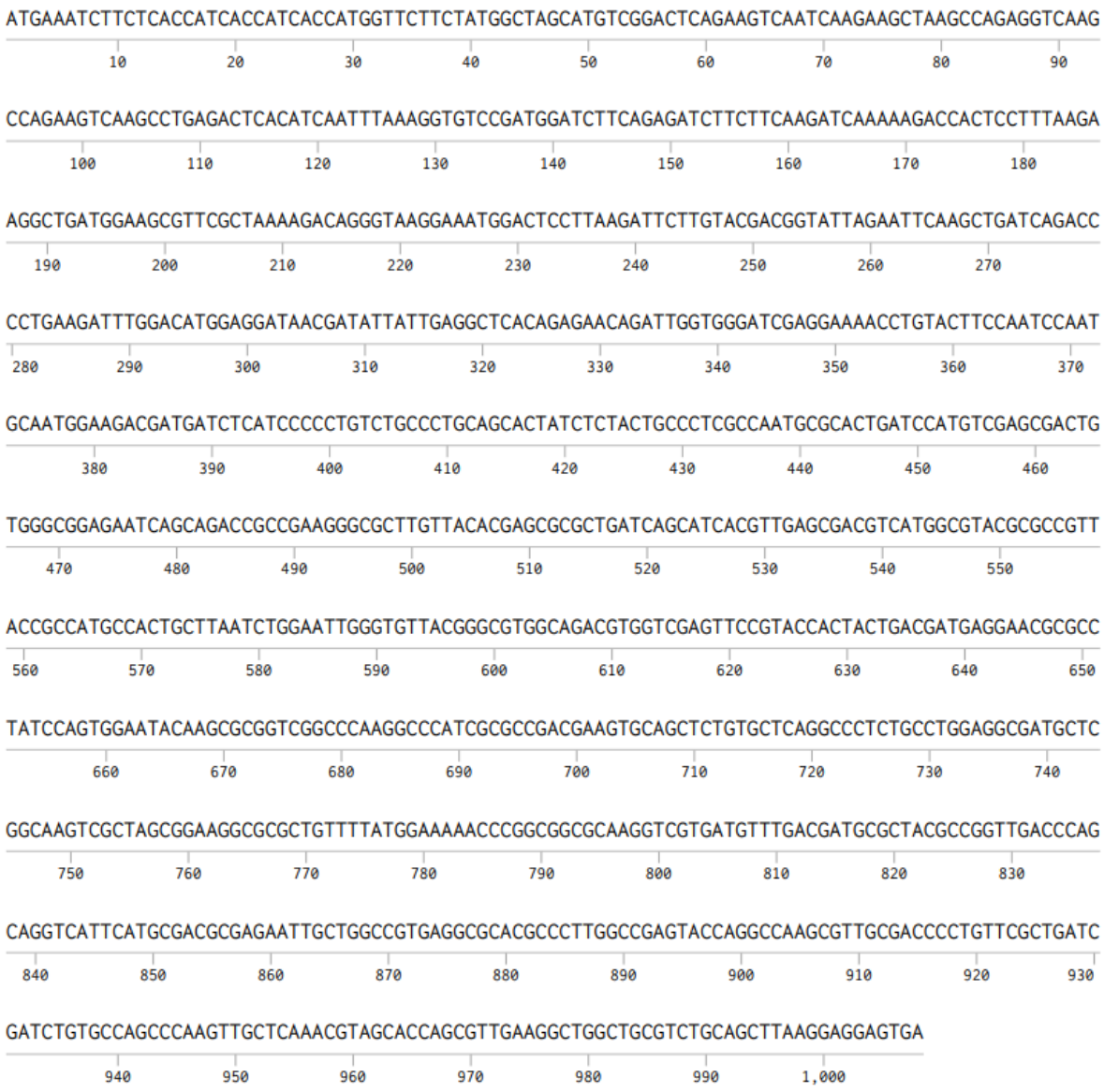

Figure 33: gBlocks sequence of Cas4 gene from type I-C CRISPR system of *Pseudomonas*.

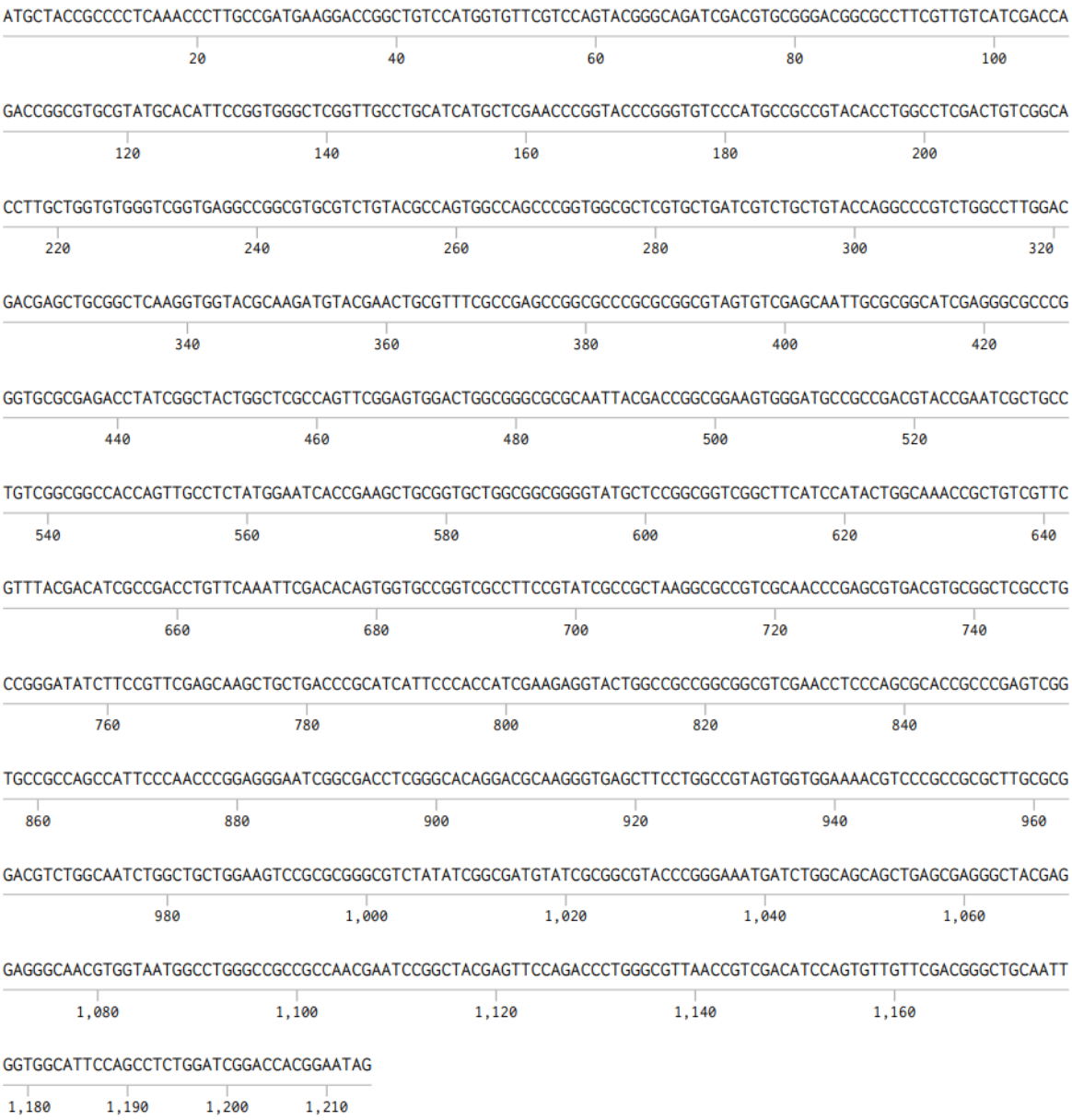

Figure 34: gBlocks sequence of Cas1-Cas2 genes from type I-E CRISPR system of *Pseudomonas.*

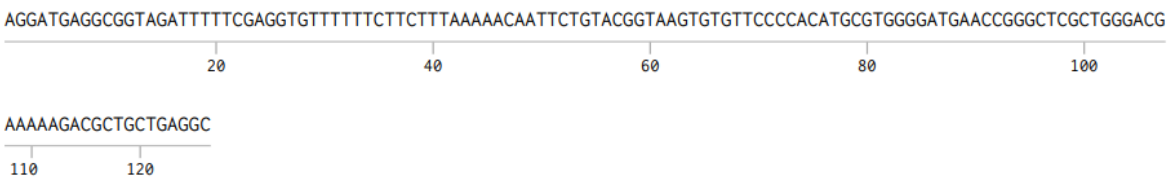

Figure 35: gBlocks sequence of Leader-Repeat genes from type I-E CRISPR system of *Pseudomonas*.

Appendix B

Primers List

Table 11: Primers used in this study.

Primer	Sequence 5'-3'	Description
T7 Fw	TAATACGACTCACTATAGGG	T7 promotor
T7 Rv	CCGCTGAGCAATAACTAGC	T7 terminator
BN082	AAGGATCCAGGAGGTATCCTCATG	Fw to detect *in vivo* acquisition in type I-E of *E.coli* (external primer)
BN172	GAAGGAGATATACATATGGCAGATCT	Rv to detect *in vivo* acquisition in type I-C of *P. aeruginosa* VA-134 (external primer binding to the backbone)
BN203	AGCGGGGATAAACCGC	Degenerate primer to detect *in vivo* acquisition in type I-E of *E.coli*
BN205	AGCGGGGATAAACCGT	Degenerate primer to detect *in vivo* acquisition in type I-E of *E.coli*
BN238	ATCGCTCAAACCACTTACGG	Rv to detect *in vivo* acquisition in type I-E of *E.coli* (external primer)
BN279	AGCGGGGATAAACCGG	Degenerate primer to detect *in vivo* acquisition in type I-E of *E.coli*
BN514	GATGGTGTCCGGGATCTC	Fw binding to pACYC
BN519	CAAAGCCCGAAAGGAAGCTGAGTT	Rv to detect *in vivo* acquisition in type I-E of *P. aeruginosa* AZPAE14509 (external primer binding to the backbone)
BN742	AAGGATCCAGGAGGTATCCTCATG	Fw to amplify KPP25 vCas4 gene from gBlocks
BN743	CTGAGTGGAAGATATGAGGTACCAA	Rv to amplify KPP25 vCas4 gene from gBlocks

Primer	Sequence 5'-3'	Description
BN744	AAGGATCCAGGAGGTGCGAC	Fw to amplify LU11 vCas4 gene from gBlocks
BN745	CGTGGTAAGACGAAATGAGGTACCAA	Rv Fw to amplify LU11 vCas4 gene from gBlocks
BN746	AAGGATCCAGGAGGGTGAGTTATG	Fw to amplify CP30A vCas4 gene from gBlocks
BN747	GCACAATGATACTAACATTTGAGGTACCAA	Rv to amplify KPP25 vCas4 gene from gBlocks
BN748	TACTTCCAATCCAATGCAATGCGCTGGTCGATCAGTAA	Fw to amplify KPP25 vCas4 gene from pTU223
BN749	GCCTCTGAGTGGAAGATATGATAACATTGGAAGTGGATAA	Rv to amplify KPP25 vCas4 gene from pTU223
BN750	TACTTCCAATCCAATGCAATGTTCAAGCAACTCGCAGG	Fw to amplify LU11 vCas4 gene from pTU224
BN751	ACTCAAGCGTGGTAAGACGAAATGATAACATTGGAAGTGGATAA	Rv Fw to amplify LU11 vCas4 gene from pTU224
BN752	TACTTCCAATCCAATGCAATGAAGTATAGGTATTCTTATTCAAGGTTA	Fw to amplify CP30A vCas4 gene from pTU225
BN753	GGAGCACAATGATACTAACATTTGATAACATTGGAAGTGGATAA	Rv to amplify CP30A vCas4 gene from pTU225
BN916	ACTATATCCGGACATCCAC	Rv to confirm vCas4 gene insertion in p2AT
BN948	GGTATCTCGACATCGACGAGA	Fw binding to KPP25 vCas4
BN984	AAGGATCCTTAGGAGGCAATGCAATGGAAGACGATGATCT	Fw to amplify Cas4 gene from pTU231 (including BamHI restriction site)
BN985	CTGCAGCTTAAGGAGGAGTGAGAATTCAA	Rv to amplify Cas4 gene from pTU231 (including EcoRI restriction site)
BN986	AAGAATTCTTAGGAGGCAATGCAATGCGGCGACAG	Fw to amplify Cas1-Cas2-Leader-Repeat fragment from pTU229 (including EcoRI restriction site)
BN987	CGGTAGCCCAGTTCCCGCAAGCTTAA	Rv to amplify Cas1-Cas2-Leader-Repeat fragment from pTU229 (including HindIII restriction site)
BN988	GCGCGTGGATTGAAACT	Degenerate primer to detect *in vivo* acquisition in type I-C of *P. aeruginosa* VA-134
BN989	GCGCGTGGATTGAAACC	Degenerate primer to detect *in vivo* acquisition in type I-C of *P. aeruginosa* VA-134
BN990	GCGCGTGGATTGAAACG	Degenerate primer to detect *in vivo* acquisition in type I-C of *P. aeruginosa* VA-134
BN991	GTAGCCCAGTTCCCGC	Rv to detect *in vivo* acquisition in type I-C of *P. aeruginosa* VA-134 (internal primer binding to Spacer1)
BN993	AAGGATCCAGGAGGTATCCTCATGCGCTGGTCGATCAG	Fw to amplify KPP25 vCas4 gene from pTU226
BN1009	TTTCCTCGGCGAGAAGGCAC	Fw to delete TTT from Cas1 gene in pTU229
BN1010	CGTGAGGAAGTACAGCATGG	Fw to delete TTT from Cas1 gene in pTU229
BN1011	AAGGATCCTTAGGAGGCAATGCAATGCTACCGCCCCTCAAA	Fw to amplify Cas1-Cas2 fragment from pTU233 (including BamHI restriction site)
BN1013	CGTGGGGATGAACCGA	Degenerate primer to detect *in vivo* acquisition in type I-E of *P. aeruginosa* AZ-PAE14509
BN1014	CGTGGGGATGAACCGT	Degenerate primer to detect *in vivo* acquisition in type I-E of *P. aeruginosa* AZ-PAE14509

Primer	Sequence 5'-3'	Description
BN1015	CGTGGGGATGAACCGC	Degenerate primer to detect *in vivo* acquisition in type I-E of *P. aeruginosa* AZ-PAE14509
BN1016	AAAAGACGCTGCTGAGGC	Rv to detect *in vivo* acquisition in type I-E of *P. aeruginosa* AZPAE14509 (internal primer binding to Spacer1)
BN1088	TCTGGATCGGACCACGGAATAGGAATTCAA	Rv to amplify Cas1-Cas2 fragment from pTU233 (including EcoRI restriction site)
BN1089	AAGAATTCAGGATGAGGCGGTAGATTTTTCGAG	Fw to amplify Leader-Repeat fragment from pTU235 (including EcoRI restriction site)
BN1090	GAAAAAGACGCTGCTGAGGCCTGCAGAA	Rv to amplify Leader-Repeat fragment from pTU235 (including PstI restriction site)
BN1127	ATGGAAGACGATGATCTC	Fw to remove His6 SUMO Tag from pTU231
BN1128	CCTGTACTTCCAATCCAATGCA	Rv to perform PCR- mediated deletion of undesirable fragment in pTU229
BN1129	ATGGCGGGCGAGATAC	Fw to perform PCR- mediated deletion of undesirable fragment in pTU229 and pTU230
BN1130	GAATTCTTAGGAGGCAATGCA	Rv to perform PCR- mediated deletion of undesirable fragment in pTU230

Appendix C

Acquisition assay in type I-E of *Pseudomonas* - Report

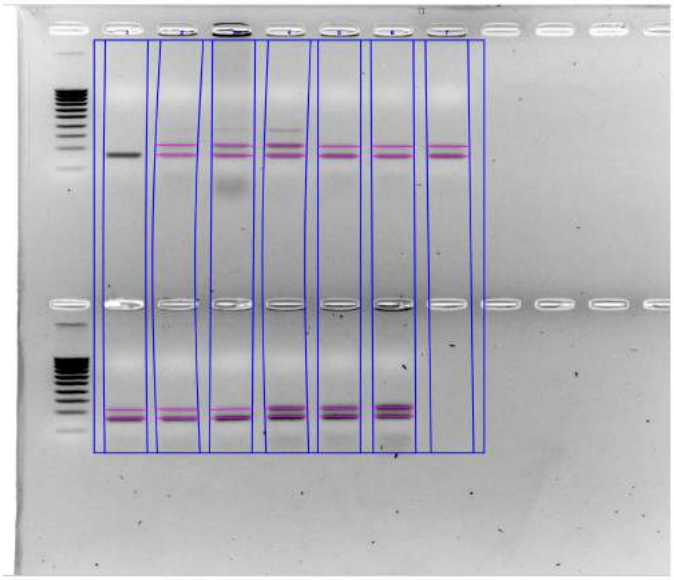

Figure 36: Agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas* The detected bands and correspondent lanes analyzed using ImageLab software are represented in blue and red, respectively.

Lane 1

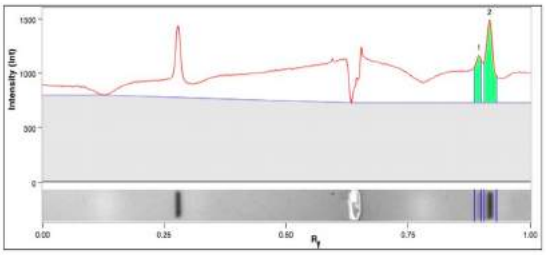

Figure 37: Lane 1 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,895	333696
2	0,916	771136

Lane 2

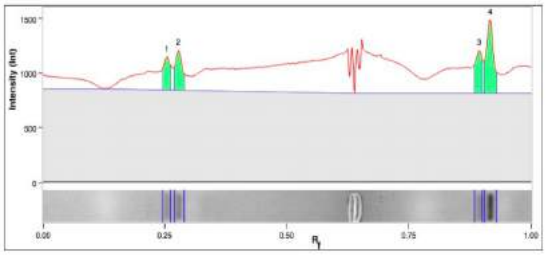

Figure 38: Lane 2 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,255	254336
2	0,279	321920
3	0,894	329600
4	0,916	644672

Lane 3

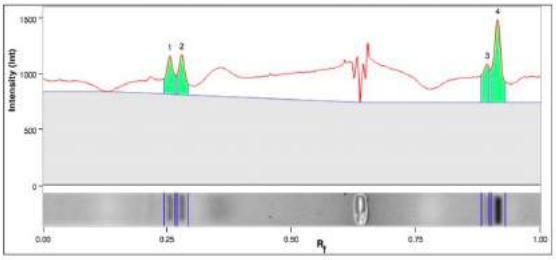

Figure 39: Lane 3 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,256	352640
2	0,280	348096
3	0,894	324992
4	0,915	812736

Lane 4

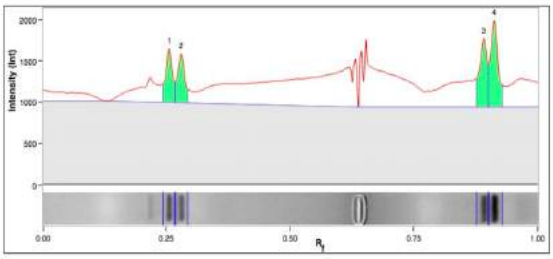

Figure 40: Lane 4 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,256	478912
2	0,280	465792
3	0,891	728512
4	0,912	882496

Lane 5

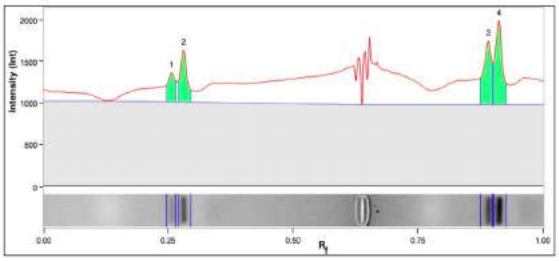

Figure 41: Lane 5 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*

Band Number	Relative Front	Volume (Int)
1	0,257	275136
2	0,281	476416
3	0,891	662848
4	0,912	802240

Lane 6

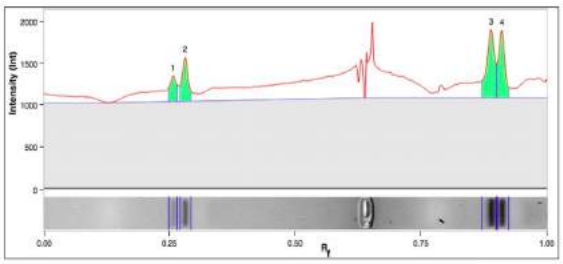

Figure 42: Lane 6 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,257	208128
2	0,281	356736
3	0,890	699520
4	0,911	566656

Lane 7

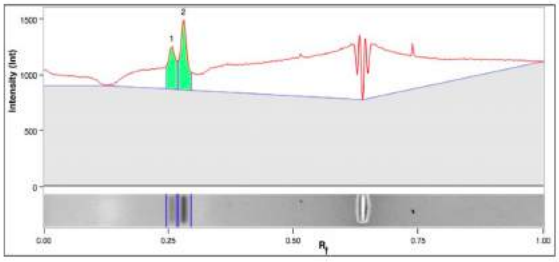

Figure 43: Lane 7 of the agarose gel obtained in the *In vivo* acquisition assays in type I-E of *Pseudomonas*.

Band Number	Relative Front	Volume (Int)
1	0,257	344448
2	0,281	532608

Appendix D

ANOVA statistical analysis

Alpha	0.05							
Dunnett's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Significant?	Summary	Adjusted P Value	D-?		
2AT vs. KPP25	0.1807	-8.655 to 9.017	No	ns	0.9999	A	KPP25	
2AT vs. LU11	10.99	2.154 to 19.83	Yes	*	0.0180	B	LU11	
2AT vs. CP30A	17.7	8.863 to 26.54	Yes	**	0.0011	C	CP30A	
Test details	Mean 1	Mean 2	Mean Diff.	SE of diff.	n1	n2	q	DF
2AT vs. KPP25	48.57	48.38	0.1807	3.068	3	3	0.05891	8
2AT vs. LU11	48.57	37.58	10.99	3.068	3	3	3.582	8
2AT vs. CP30A	48.57	30.87	17.7	3.068	3	3	5.768	8

Figure 44: Ordinary one-way ANOVA multiple comparison statistical test results.