Cas4 solo in phages enhances host CRISPR autoimmunity

Cátia Sofia Marques Pereira

catia.m.pereira@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal ; Technische Universiteit Delft, Delft, Netherlands

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

Introduction

Dynamic interactions between phages and hosts have been studied since the early days of molecular biology. This phage-host arms race shaped the evolution of microbes to evolve a vast diversification of defense mechanisms that can be classified as innate or adaptive immune systems. CRISPR and their associated *cas* genes encode one such adaptive immune system mechanism [12].

Found in approximately 45% of bacteria and 85% of *archaea* [15], the CRISPR array consist of a cluster of a highly variable number of repeats, interspaced by spacers and a *leader* sequence [20]. 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 [13]. The CRISPR-Cas systems have evolved a vast diversification being categorized into two classes, six types and more than twenty subtypes [15].

CRISPR-Cas Mechanism. 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 1). 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.

To avoid acquisition of spacers from the host DNA that would lead to autoimmunity, the CRISPR systems uses the DNA machinery repair of the hosts to generate protospacers [20].

Cas1 is the most conserved Cas protein, present in all of the six types of CRISPR systems [15]. In the context of CRISPR immunity, this protein interacts with Cas2 forming a complex responsible for spacer integration (Cas1-Cas2) [33]. This complex has two separate DNA-binding proteins mediating the connection between the incoming protospacer and the CRISPR array [20]. The new spacers acquired by the Cas1-Cas2 complex are predominately incorporated in the leader end of the CRISPR array [33]. 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 [33].

The identification of convenient protospacers is based on the presence of a protospacer-adjacent motif (PAM) [21] that ensures the correct orientation of protospacers in the CRISPR array [31].

Other Cas protein that is known to be implicated in the adaptation phase in several subtypes of CRISPR-Cas systems is the Cas4 protein. In the CRISPR loci the *cas4* genes were found either adjacent to *cas1* and *cas2* or, in some cases, fused with *cas1* [15]. 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 [17, 14, 27, 34, 29].

The adaptation stage of CRISPR-Cas mechanism is followed by the expression phase in which the CRISPR array and CRISPR locus are transcribed [4]. The transcription of the CRISPR array leads to the formation of a long precursor CRISPR-RNA molecule (pre-crRNA) [4] which is further processed forming hairpin-like structures that later produce mature crRNAs [19].

Interference is the last stage of the CRISPR-Cas immune system mechanism (Figure 1). 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 [19]. Once generated, crRNAs use their base-pairing potential and serve as guides for the recognition of invasive targets [19]. After PAM scanning and recognition, in the cases where the crRNA fully matches the target

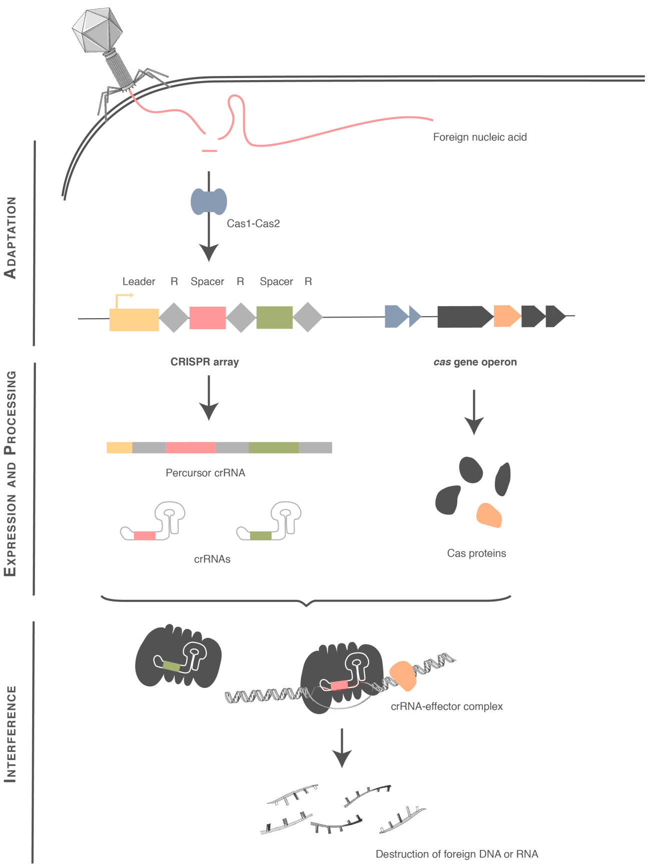


Figure 1: Schematic representation of CRISPR-Cas immune system mechanism During adaptation, the Cas1-Cas2 complex select and incorporate new spacers in the CRISPR array. During expression and processing, the CRISPR array is transcribed to produce crRNA that forms the crRNA-effector complex by binding Cas proteins. During interference, the foreign DNA is recognized and degraded. Modified from Jackson *et al.*, 2017.

nucleic acid, nuclease Cas proteins are recruited to the side promoting the final destruction of the invaders [32].

Cas4 solo. It was found that the *cas4* gene can also be found not associated with the CRISPR-cas loci (sCas4) in some prokaryotic genomes as well as plasmids and bacteriophages (vCas4). Recently, a phylogenomic study of Cas4 family nucleases was performed allowing the detection of 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 [11] confirming the diversity of these solo Cas4 proteins.

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. Regarding the influence of vCas4 in the CRISPR system, in the case they do interact, it can happen 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.

Materials and Methods

Bacterial Strains and Growth Conditions. The bacterial strains used in this study were *E. coli* DH5α and *E. coli* BL21-AI [3, 28]. All bacterial cultures were grown in LB media at 37C and continuous shaking at 180 rpm or in LBA plates. When required, antibiotics and inducers were supplied.

gBlocks. Some DNA fragments used in this study were chemically synthesized and ordered as gBlocks Gene Fragments as in the Cas4 homolog genes encoded in LU11 and KPP25 *Pseudomonas* phage genome, CP30A *Campylobacter* phage genome and the *Pseudomonas* type I-C *Cas1-Cas2-Leader-Repeat* and *cas4* genes and the *Pseudomonas* type I-E *cas1-cas2* and *Leader-Repeat* genes.

Plasmids. The plasmids used in this study and corresponding selection markers are described in Supplementary Data.

Polymerase Chain Reactions. Two different PCR were performed, either using Q5 DNA Polymerase or OneTaq DNA Polymerase. The components used in the mixture, their final volumes and the PCR program applied were dependent on the polymerase used.

Restriction Enzyme Cloning. In Restriction Enzyme Cloning, the DNA fragments were digested using appropriate restriction enzymes and subsequently ligated. For digestion, the CutSmart buffer 10x was mixed with the correct enzymes, the product to digest and the volume of Milli-Q water necessary to obtain the final desired volume and incubated at 4 C overnight. After that, restriction enzymes were inactivated by heat or by DNA purification. The restriction enzymes used in this study were EcoRI-HF, BamHI-HF, HindIII-HF, PstI-HF and KpnI-HF. The amount of each fragment needed to obtain the final molar ration of 1:3 (vector:insert) was mixed with T4 Ligase buffer 10x, T4 ligase and Milli-Q water and incubated at 4 C overnight.

Ligation-independent Cloning. All the plasmids in which the His6 SUMO Tag was inserted were obtained by LIC. In this cloning process, the fragments to insert were amplified by Q5 DNA polymerase PCR and the vector (p13SS) was linearized and after purification of both PCR amplicon and linearized vector, LIC reactions were prepared and incubated at 22 C for 30 minutes then 75 C for 20 minutes. Finally, both the LICed PCR and LICed vector were combined and incubated at room temperature during 10 minutes.

PCR-mediated deletion. The PCR-mediated deletion process was implemented to delete a mutation, to remove an additional 80bp sequence from the gBlocks insertion in two plasmids and also to remove the fragment codifying for the His6 SUMO Tag. To complete the deletions, first, a Q5 DNA polymerase PCR using primers flanking the region to delete was performed and after DNA purification, 1 L of dpnI was added to the PCR product and this mixture was incubated at 4 C overnight. This enzyme was inactivated by DNA purification and the 5' phosphorylation reaction was prepared by addition of T4 Polynucleotide Kinase. After 45 minutes of incubation at 37 C, T4 DNA ligase was added and the final mixture was incubated at 4 C overnight.

pGEM-T Vector System. The pGEM-T Vector System was used to clone PCR products for White-Blue Screening. The ligation reactions were performed according to manufacturer indications.

DNA Purification. Plasmid extraction from bacterial cultures was done using GeneJET Plasmid Miniprep Kit. Purification of PCR products and DNA cleaningwas performed using GeneJET PCR Purification Kit. Both kits were used as per manufacturer instructions.

Sequencing. DNA sequencing was done by Sanger method and outsourced to Macrogen Inc. Amsterdam.

Transformation. 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). 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 and from this culture, chemical competent cells were prepared using the *Mix & Go* E. coli Transformation Kit (Zymo Research) according to manufacturer instructions.

***In vivo* spacer acquisition assays.**To detect acquisition in both types I-E and I-C CRISPR systems of *P. aeruginosa* were transformed all the plasmids carrying the CRISPR machinery components. The transformed cells were induced by supplementation of L-arabinose and IPTG and grown at 37 C overnight. Spacer acquisition was monitored by PCR. 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^{TM} Software report tool. With the values obtained was further quantified the normalized percentage of expanded band and this analysis was further complemented with an analysis of variance (ANOVA). These assays were followed by the separation of the PCR products corresponding to the expanded CRISPR array from the parental ones using the BluePippin automated agarose-electrophoresis system as per manufacturer instructions. 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.

Expanded CRISPR array sequencing and protospacer analysis. The expanded CRISPR array collected from BluePippin was inserted in the pGEM-T vector and transformed in *E. coli* DH5α cells for Blue-White Screening. The sequencing of the white colonies was done by Sanger method and outsourced to GATC (Eurofins Genomics). The sequencing results were further analysed 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.

Protein expression and purification by Ni-NTA Affinity Chromatography. To overexpress 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 37°C 180 rpm, 2L of LB media were inoculated and later induced by supplementation of L-arabinose and IPTG and grown at 20 C and continuous shaking at 180 rpm overnight. Cells were then harvested by centrifugation, ressuspended in chilled Lysis Buffer and cOmplete EDTA-free Protease Inhibitor Cocktai and, later, lysed by French Press (1000 bar). The lysate was cleared by centrifugation and filtered through a 0,45mm syringe filter. The HIS-Select Nickel Affinity Gel was washed with chilled Lysis Buffer,incubated with the clarified lysate and then load in a gravity disposable column. Visualization of the purified proteins was done using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

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. After centrifugation, the supernatant was applied to a Superdex 200 10/300 GL column connected to an ÄKTA purifier system. The sample was eluted with Elution Buffer and detected at 260 and 400 nm and the fractions of interest were collected. Visualization of the purified proteins was done using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

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

***In vitro* Nuclease Activity Assays.** 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), respectively. 500 nmol of the purified protein were mixed with 100 ng of DNA substrate in 10x reaction buffer (supplemented with $MgCl_2$ or $MnCl_2$) and Milli-Q water. In the case of KPP25 vCas4, this mixture was incubated at 37°C during 2 hours. With samples on ice, the reaction was quenched by the addition of Proteinase K. To evaluate the activity, samples were mixed with 6X Loading dye, visualized in a 1% agarose gel and stained with SYBR Gold. 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.

Exonuclease Activity Assay. 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 95C during 10 minutes. Samples were further separated on a PAGE-denaturing gel (20% polyacrylamide, 7M urea, 1 TBE).

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 pluginand 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.

Results

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. From this analysis were retrieved 112 sequences in total that were found to be encoded in a high diversity of phages. To perform phylogenomic studies, these proteins were aligned by MAFFT alignment [8] 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) [16]. This resulted in the inactivation of the nuclease activity of the Cas4 protein by the lost of some of the catalytic amino acid residues [16]. 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 (Supplementary Data) it was possible to see that 4 cysteine residues are very well conserved in all the vCas4 proteins in study, including the TTV1 protein and the Cas4 proteins encoded in the CRISPR loci. This four cysteines are known to be highly conserved in Cas4 proteins which are presumably responsible for the coordination of the iron-sulfur cluster [34]. The fact they are conserved in almost all the proteins presented in the alignment shows high evidence that these proteins might have similar function. However, structural studies of Cas4 have shown that not only the 4 cysteines are conserved but also some other residues[18], including a lysine and an additional aspartic acid residue [34]. It is possible to see that, as the four cysteine residues, the aspartic acid residue is very well conserved. However, 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 except the TTV1 Cas4 protein. 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.

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 studying 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.

Contrarily to supposed, the TTV1 protein clustered together with other vCas4 proteins and not independently in the phylogenimc tree obtained (Supplementary Data). In this case, if the vCas4 proteins were analyzed only by the phylogenomic tree obtained, this protein with lost functionality wouldn't be identified. Moreover, the Cas4 proteins known to be associated with type I CRISPR-Cas systems that were included in this study are not clustered together but placed in different regions of the phylogenomic tree. 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. 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. The protein alignment was then used to choose proteins to perform further experimental assays. The first criteria of choice was the conservation of the four cysteines and the one lysine domains. The proteins chosen were the ones encoded in *Campylobacter* phage CP30A, *Pseudomonas* phage KPP25 and *Pseudomonas* phage LU11.

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 and 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 are accessible and well studied. No CRISPR-Cas system have ever been identified in *Pseudomonas putida* infected by LU11 phage, however, in *Pseudomonas aeruginosa* three different CRISPR-Cas systems were identified [5]. 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.

Since it was important to make sure that the vCas4 proteins in study were not related with the TTV1 protein, were 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 [6] according to genes' function. 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. With this analysis it was 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. It is then 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.

***In vivo* Acquisition Assays.** 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. This way, by performing the PCR to detect acquisition, two different bands, corresponding to the amplification of two different populations, are obtained.

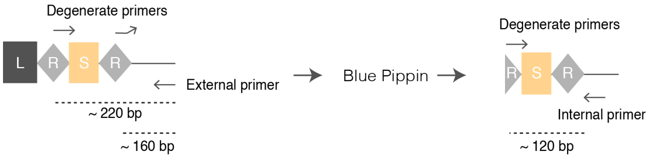


Figure 2: Binding sites of the primers used in the *in vivo* acquisition assays and expected sizes of the expanded and non-expanded bands. In the right, the degenerate forward primers bind to the *repeat* fragment (R) and the external reverse primer binds to the backbone. Are also represented the *leader* (L) and *spacer* (S) fragments. In the left 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.

In all samples (except the negative control) it was possible to detect a clear expanded band (+1 band) meaning that new spacers were acquired by the CRISPR-Cas system.

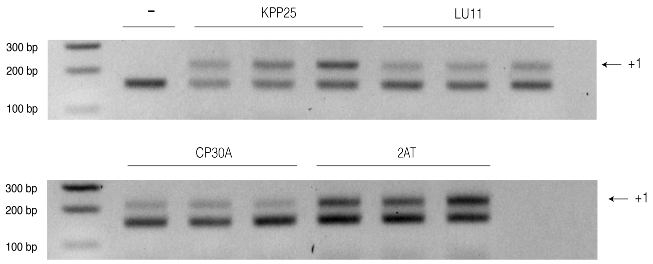


Figure 3: PCR of the *in vivo* acquisition assays in the type I-E of *Pseudomonas* CRISPR-Cas system. Top (right to left): Negative control, KPP25 vCas4 and LU11 vCas4; Down (right to left): CP30A and 2AT empty plasmid (positive control). The band corresponding to the amplification of the expanded (+1) population in which CRISPR array the new spacers were incorporated is marked with a black arrow.

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 4).

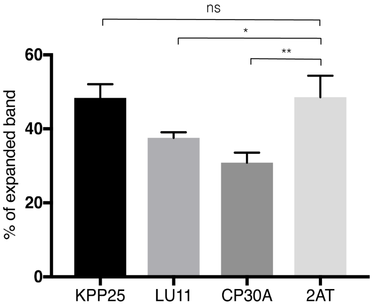


Figure 4: Percentage of the expanded band in the PCR of the *in vivo* acquisition assays in the type I-E of *Pseudomonas* CRISPR-Cas system. Percentages obtained by division of the intensities of the expanded band divided by the sum of the intensities of both expanded and non expanded band for the cases 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. 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 1).

Table 1: 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.

	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

In the cases that KPP25 vCas4, LU11 vCas4 and no vCas4 protein was present (p2AT 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 1). 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.

The length of the newly acquired spacers and their consensus PAM was also determined (Figure 5).

It was possible to conclude that the most predominant spacer length of the novel spacers acquired was 33 bp and their consensus PAM was 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* [25], we can conclude that the vCas4 of this study does not have influence in those parameters.

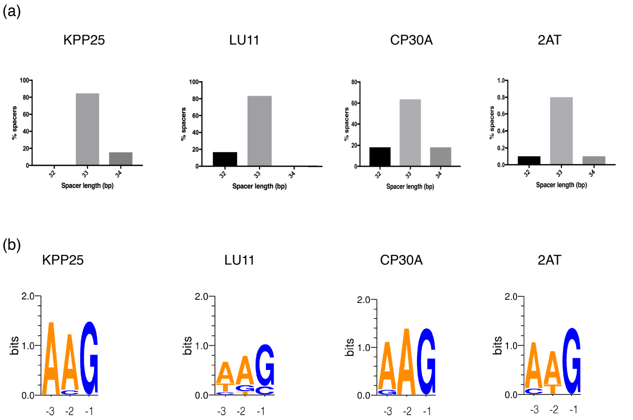


Figure 5: Length and PAM consensus sequence of the novel spacers acquired by the CRISPR array of type I-E of *Pseudomonas* (a) Spacer length distribution in the presence of KPP25, LU11 and CP30A vCas4 and in its absence (2AT empty plasmid). (b) PAM consensus sequence of the novel spacers acquired by 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.

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.

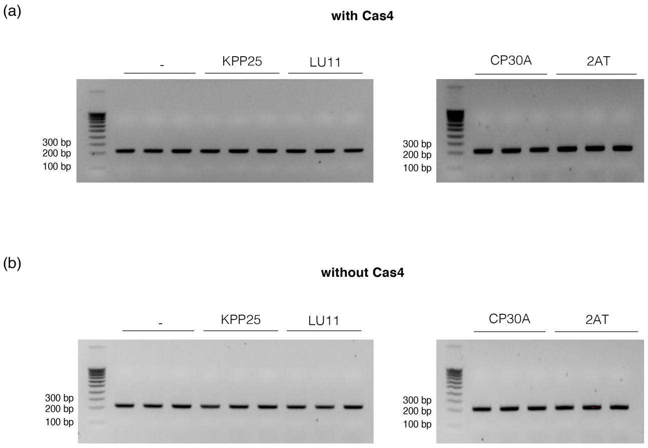


Figure 6: PCR of the *in vivo* acquisition assays in the type I-C of *Pseudomonas* CRISPR-Cas system in the (a) presence and (b) absence of the Cas4 protein from type I-C of *Pseudomonas*. Right to left: Negative control, KPP25 vCas4, LU11 vCas4, CP30A vCas4 and 2AT empty plasmid (positive control). Only the band corresponding to the amplification of the non expanded (+0) population, in which no new spacers were incorporated, is present.

After performing the acquisition 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. 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. By analysis of the results obtained it is possible to see that, in fact, no acquisition can be detected since no amplification was obtained in the PCR.

Finally, it allows the conclusion that the type I-C of *Pseudomonas* is not able to acquire novel spacers (naive acquisition). Only *in vitro* naive acquisition was previously described in the type I-C CRISPR-Cas systems [17] and no *in vivo* acquisition was ever described. The only *in vivo* acquisition described in this type of CRISPR systems was priming acquisition [25]. 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]. This process by which pre-existing spacers facilitate rapid spacer acquisition is known as primed spacer acquisition (or priming) [7, 31]. Unlike naïve 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 [7]. 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*.

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, further co-purification assays were then performed in order to understand if this protein strongly interacts with the Cas1-Cas2 complex. 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.

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.

By analysis of the obtained results it is possible to see that in the case of KPP25 vCas4 protein (Figure 7 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 7 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.

These samples were additionally evaluated by 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. 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.

In the co-purification of the type I-C, for the KPP25 vCas4 protein (Figure 8) 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.

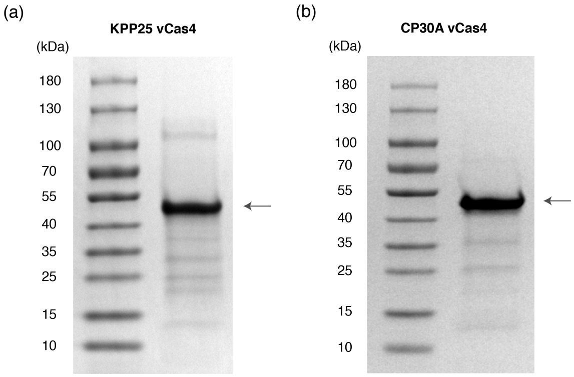


Figure 7: SDS-page gel resulting from the assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-E of *Pseudomonas* in the case of (a) KPP25 vCas4 or (b) CP30A vCas4. The band demonstrating the positive purification of KPP25 vCas4 (43 kDa) and of CP30A vCas4 (44 kDa) is marked with 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.

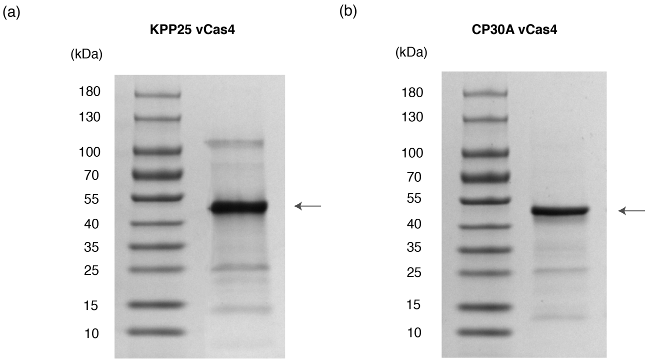


Figure 8: SDS-page gel resulting from the assays to evaluate vCas4 interaction with the Cas1-Cas2 complex of type I-C of *Pseudomonas* in the case of (a) KPP25 vCas4 or (b) CP30A vCas4. The band demonstrating the positive purification of KPP25 vCas4, with approximately 43 kDa is marked with a black arrow 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.

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. This way, the all the vCas4 proteins were first purified using Ni-NTA Affinity Chromatography and further subjected to an additional size-exclusion chromatography. After size exclusion, 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.

In the first assay, these proteins were incubated with linear double-stranded DNA, circular double-stranded DNA and circular single-stranded DNA during two hours at 37°C in the presence of two different buffers (MgCl$_2$ and MnCL$_2$).

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 9). 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 MnCl2 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 9). This result suggests that this protein is actively degradating the DNA sample into single nucleotides that can't be stained using the DNA loading dye. In order to confirm this hypothesis, the same *in vitro* activity assay were repeated over time and with additional supplementation of EDTA that was used in this assay with the objective of stopping the degradation reactions [24].

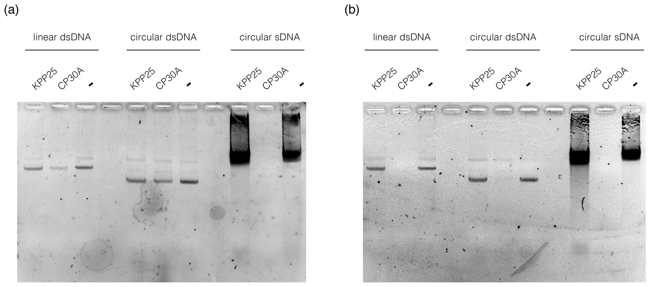


Figure 9: *In vitro* assay for determination of vCas4 protein activity. Both KPP25 and CP30A vCas4 proteins were incubated with different types of DNA (linear double-stranded DNA, circular double-stranded DNA and circular single-stranded DNA) during two hours at 37^{o}C in the presence of (a) MgCl$_2$ buffer and (b) MnCl$_2$ buffer.

The reactions immediately after incubation and after 10 and 30 of reaction, for all the DNA samples tested before. 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).

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. The observed 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.

Discussion

The Cas4 proteins has long been implicated in immune adaptation, forming complex with Cas1 [17] and selecting and orienting PAM-compatible spacers. [14, 29]. 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 [11].

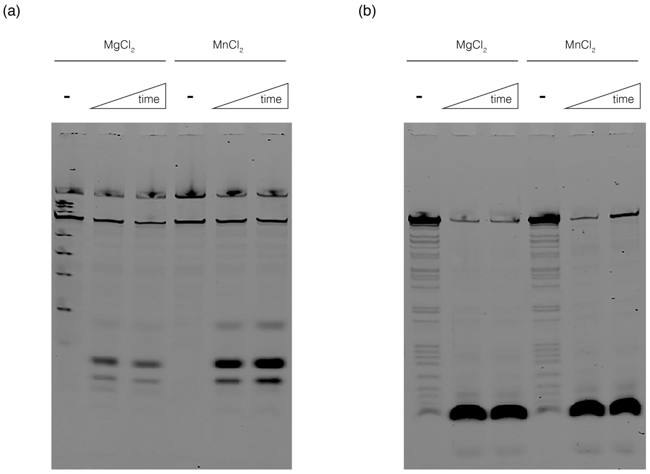


Figure 10: 20% SDS-page gel to detect exonuclease activity of CP30A vCas4 over (a) 5'-end labelled oligonucleotides or (b) 3'-end labelled oligonucleotides. Both reactions were performed in the presence of both MnCl$_2$ and MgCl$_2$ buffers and over time (10 and 30 minutes). After incubation, the reactions were stopped by EDTA supplementation.

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 [16]. 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[10, 9]. 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 [5]. 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 [26, 1]. 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 [27] and this processing ensures the formation of optimal protospacers [17]. 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.

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 emergence of pathogenic bacteria resistant to most, if not all, of the antimicrobial agents available has become a critical problem in modern medicine [30, 23]. 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 [30].

Even if phage therapy might look like a promising alternative to the use of antibiotics, the evolution of bacterial resistance to a particular phage, just as to an antibiotic, is inevitable [22]. 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.

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.

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. 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.

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 instead of magnesium ions. 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.

References

[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] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] Anders S. Nilsson. Phage therapy-constraints and possibilities, 2014.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

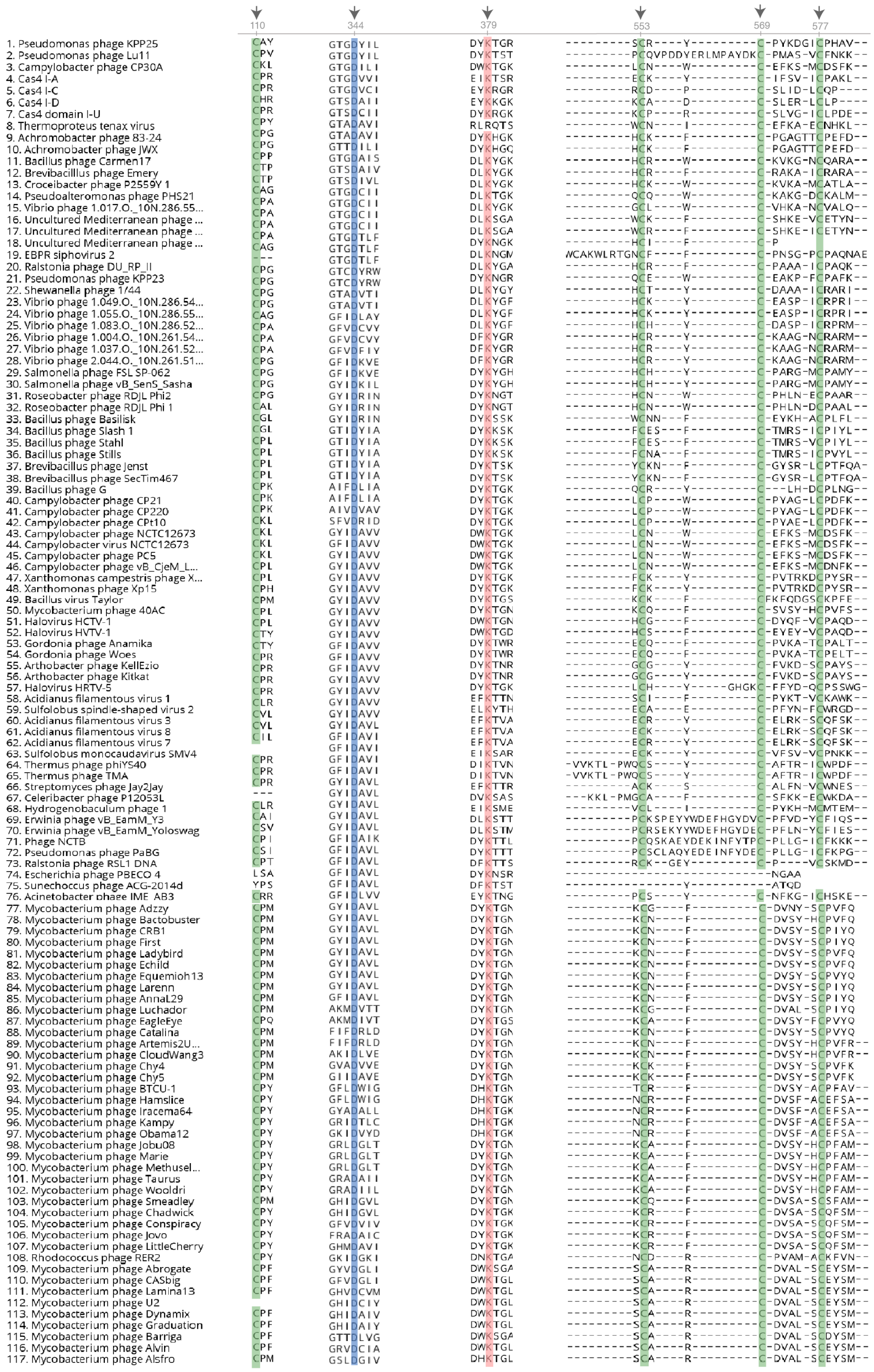
[33] Ido Yosef, Moran G Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research*, 40(12):5569–76, 2012.

[34] 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.

SUPPLEMENTARY DATA A

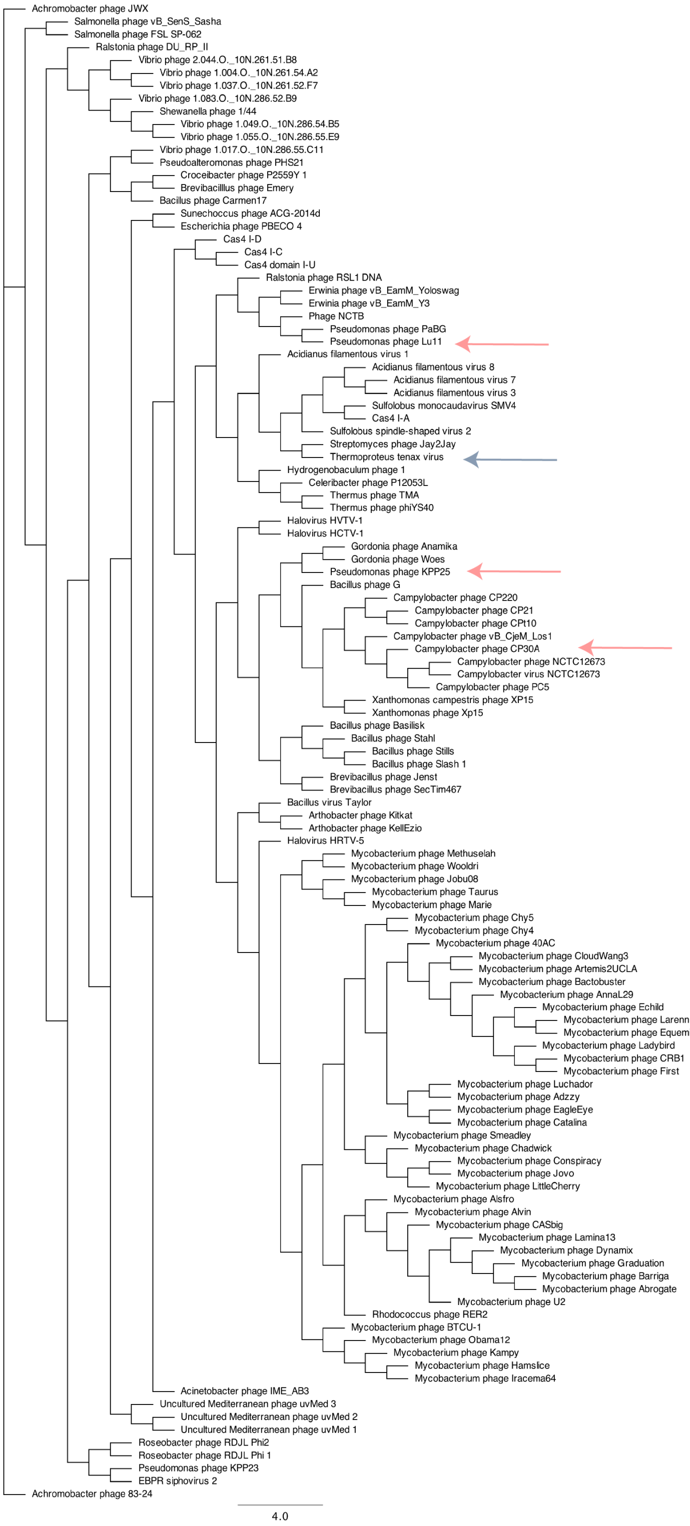
Plasmid	Description	Resistance	Reference
p2AT	pET LIC cloning (2A-T)	Amp	Addgene # 29665
p13SS	pET His6 Sumo TEV cloning (13S-S)	Spec	Addgene # 48329
pACYC	pACYCDuet-1	Cm	Novagen # 71147
pCas12	pACYC with *E. coli* K-12 type I-E Cas1-Cas2	Cm	Not published
pTU223	p2AT with KPP25 vCas4	Amp	This study
pTU224	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

SUPPLEMENTARY DATA B


110
344
379
553
569
577
1. Pseudomonas phage KPP25
2. Pseudomonas phage Lu11
3. Campylobacter phage CP30A
4. Cas4 I-A
5. Cas4 I-C
6. Cas4 I-D
7. Cas4 domain I-U
8. Thermoproteus tenax virus
9. Achromobacter phage 83-24
10. Achromobacter phage JWX
11. Bacillus phage Carmen17
12. Brevibacilllus phage Emery
13. Croceibacter phage P2559Y 1
14. Pseudoalteromonas phage PHS21
15. Vibrio phage 1.017.O._10N.286.55...
16. Uncultured Mediterranean phage ...
17. Uncultured Mediterranean phage ...
18. Uncultured Mediterranean phage ...
19. EBPR siphovirus 2
20. Ralstonia phage DU_RP_II
21. Pseudomonas phage KPP23
22. Shewanella phage 1/44
23. Vibrio phage 1.049.O._10N.286.54...
24. Vibrio phage 1.055.O._10N.286.55...
25. Vibrio phage 1.083.O._10N.286.52...
26. Vibrio phage 1.004.O._10N.261.54...
27. Vibrio phage 1.037.O._10N.261.52...
28. Vibrio phage 2.044.O._10N.261.51...
29. Salmonella phage FSL SP-062
30. Salmonella phage vB_SenS_Sasha
31. Roseobacter phage RDJL Phi2
32. Roseobacter phage RDJL Phi 1
33. Bacillus phage Basilisk
34. Bacillus phage Slash 1
35. Bacillus phage Stahl
36. Bacillus phage Stills
37. Brevibacillus phage Jenst
38. Brevibacillus phage SecTim467
39. Bacillus phage G
40. Campylobacter phage CP21
41. Campylobacter phage CP220
42. Campylobacter phage CPt10
43. Campylobacter phage NCTC12673
44. Campylobacter virus NCTC12673
45. Campylobacter phage PC5
46. Campylobacter phage vB_CjeM_L...
47. Xanthomonas campestris phage X...
48. Xanthomonas phage Xp15
49. Bacillus virus Taylor
50. Mycobacterium phage 40AC
51. Halovirus HCTV-1
52. Halovirus HVTV-1
53. Gordonia phage Anamika
54. Gordonia phage Woes
55. Arthobacter phage KellEzio
56. Arthobacter phage Kitkat
57. Halovirus HRTV-5
58. Acidianus filamentous virus 1
59. Sulfolobus spindle-shaped virus 2
60. Acidianus filamentous virus 3
61. Acidianus filamentous virus 8
62. Acidianus filamentous virus 7
63. Sulfolobus monocaudavirus SMV4
64. Thermus phage phiYS40
65. Thermus phage TMA
66. Streptomyces phage Jay2Jay
67. Celeribacter phage P12053L
68. Hydrogenobaculum phage 1
69. Erwinia phage vB_EamM_Y3
70. Erwinia phage vB_EamM_Yoloswag
71. Phage NCTB
72. Pseudomonas phage PaBG
73. Ralstonia phage RSL1 DNA
74. Escherichia phage PBECO 4
75. Synechococcus phage ACG-2014d
76. Acinetobacter phage IME_AB3
77. Mycobacterium phage Adzzy
78. Mycobacterium phage Bactobuster
79. Mycobacterium phage CRB1
80. Mycobacterium phage First
81. Mycobacterium phage Ladybird
82. Mycobacterium phage Echild
83. Mycobacterium phage Equemioh13
84. Mycobacterium phage Larenn
85. Mycobacterium phage AnnaL29
86. Mycobacterium phage Luchador
87. Mycobacterium phage EagleEye
88. Mycobacterium phage Catalina
89. Mycobacterium phage Artemis2U...
90. Mycobacterium phage CloudWang3
91. Mycobacterium phage Chy4
92. Mycobacterium phage Chy5
93. Mycobacterium phage BTCU-1
94. Mycobacterium phage Hamslice
95. Mycobacterium phage Iracema64
96. Mycobacterium phage Kampy
97. Mycobacterium phage Obama12
98. Mycobacterium phage Jobu08
99. Mycobacterium phage Marie
100. Mycobacterium phage Methusel...
101. Mycobacterium phage Taurus
102. Mycobacterium phage Wooldri
103. Mycobacterium phage Smeadley
104. Mycobacterium phage Chadwick
105. Mycobacterium phage Conspiracy
106. Mycobacterium phage Jovo
107. Mycobacterium phage LittleCherry
108. Rhodococcus phage RER2
109. Mycobacterium phage Abrogate
110. Mycobacterium phage CASbig
111. Mycobacterium phage Lamina13
112. Mycobacterium phage U2
113. Mycobacterium phage Dynamix
114. Mycobacterium phage Graduation
115. Mycobacterium phage Barriga
116. Mycobacterium phage Alvin
117. Mycobacterium phage Alsfro

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).

SUPPLEMENTARY DATA C


Achromobacter phage JWX
Salmonella phage vB_SenS_Sasha
Salmonella phage FSL SP-062
Ralstonia phage DU_RP_II
Vibrio phage 2.044.O._10N.261.51.B8
Vibrio phage 1.004.O._10N.261.54.A2
Vibrio phage 1.037.O._10N.261.52.F7
Vibrio phage 1.083.O._10N.286.52.B9
Shewanella phage 1/44
Vibrio phage 1.049.O._10N.286.54.B5
Vibrio phage 1.055.O._10N.286.55.E9
Vibrio phage 1.017.O._10N.286.55.C11
Pseudoalteromonas phage PHS21
Croceibacter phage P2559Y 1
Brevibacilllus phage Emery
Bacillus phage Carmen17
Sunechoccus phage ACG-2014d
Escherichia phage PBECO 4
Cas4 I-D
Cas4 I-C
Cas4 domain I-U
Ralstonia phage RSL1 DNA
Erwinia phage vB_EamM_Yoloswag
Erwinia phage vB_EamM_Y3
Phage NCTB
Pseudomonas phage PaBG
Pseudomonas phage Lu11
Acidianus filamentous virus 1
Acidianus filamentous virus 8
Acidianus filamentous virus 7
Acidianus filamentous virus 3
Sulfolobus monocaudavirus SMV4
Cas4 I-A
Sulfolobus spindle-shaped virus 2
Streptomyces phage Jay2Jay
Thermoproteus tenax virus
Hydrogenobaculum phage 1
Celeribacter phage P12053L
Thermus phage TMA
Thermus phage phiYS40
Halovirus HVTV-1
Halovirus HCTV-1
Gordonia phage Anamika
Gordonia phage Woes
Pseudomonas phage KPP25
Bacillus phage G
Campylobacter phage CP220
Campylobacter phage CP21
Campylobacter phage CPt10
Campylobacter phage vB_CjeM_Los1
Campylobacter phage CP30A
Campylobacter phage NCTC12673
Campylobacter virus NCTC12673
Campylobacter phage PC5
Xanthomonas campestris phage XP15
Xanthomonas phage Xp15
Bacillus phage Basilisk
Bacillus phage Stahl
Bacillus phage Stills
Bacillus phage Slash 1
Brevibacillus phage Jenst
Brevibacillus phage SecTim467
Bacillus virus Taylor
Arthobacter phage Kitkat
Arthobacter phage KellEzio
Halovirus HRTV-5
Mycobacterium phage Methuselah
Mycobacterium phage Wooldri
Mycobacterium phage Jobu08
Mycobacterium phage Taurus
Mycobacterium phage Marie
Mycobacterium phage Chy5
Mycobacterium phage Chy4
Mycobacterium phage 40AC
Mycobacterium phage CloudWang3
Mycobacterium phage Artemis2UCLA
Mycobacterium phage Bactobuster
Mycobacterium phage AnnaL29
Mycobacterium phage Echild
Mycobacterium phage Larenn
Mycobacterium phage Equemioh13
Mycobacterium phage Ladybird
Mycobacterium phage CRB1
Mycobacterium phage First
Mycobacterium phage Luchador
Mycobacterium phage Adzzy
Mycobacterium phage EagleEye
Mycobacterium phage Catalina
Mycobacterium phage Smeadley
Mycobacterium phage Chadwick
Mycobacterium phage Conspiracy
Mycobacterium phage Jovo
Mycobacterium phage LittleCherry
Mycobacterium phage Alsfro
Mycobacterium phage Alvin
Mycobacterium phage CASbig
Mycobacterium phage Lamina13
Mycobacterium phage Dynamix
Mycobacterium phage Graduation
Mycobacterium phage Barriga
Mycobacterium phage Abrogate
Mycobacterium phage U2
Rhodococcus phage RER2
Mycobacterium phage BTCU-1
Mycobacterium phage Obama12
Mycobacterium phage Kampy
Mycobacterium phage Hamslice
Mycobacterium phage Iracema64
Acinetobacter phage IME_AB3
Uncultured Mediterranean phage uvMed 3
Uncultured Mediterranean phage uvMed 2
Uncultured Mediterranean phage uvMed 1
Roseobacter phage RDJL Phi2
Roseobacter phage RDJL Phi 1
Pseudomonas phage KPP23
EBPR siphovirus 2
Achromobacter phage 83-24
4.0

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.