Molecular evolution of the anti-CRISPR family AcrIF7 reveals conservation despite mutation tolerance

Wendy Figueroa[1†*], Adrián Cazares[2,3†*], Daniel Cazares[4], Ana de la Cruz[5], Luis Kameyama[5], Gabriel Guarneros[5*]

[1] Department of Biochemistry, University of Cambridge, Cambridge, Cambridgeshire, CB2 1QW, United Kingdom

[2] EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

[3] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom

[4] Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, 62210, Mexico

[5] Department of Genetics and Molecular Biology, Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico City, 07360, Mexico

†These authors contributed equally

* To whom correspondence should be addressed. Email: wf255@cam.ac.uk

Correspondence may also be addressed to:

acaza@ebi.ac.uk

gguarner@cinvestav.mx

ABSTRACT

Anti-CRISPRs are molecules typically encoded in mobile elements and capable of blocking CRISPR-Cas systems. Since their discovery, numerous anti-CRISPR families have been identified. However, the spread and sequence diversity of members within a family, and the implications of the underlying patterns of dissemination and genetic variation for the evolution and function of the anti-CRISPR, have not been investigated. Here, we use AcrIF7 as a model to explore the distribution and molecular evolution of an anti-CRISPR family. Our comparative analyses uncovered five sub-clusters and prevalent sequence variants within the group. AcrIF7 homologs display high similarity, particularly within sub-clusters, and are mainly associated with *Pseudomonas aeruginosa*. Still, *acrIF7* genes were identified in distinct genetic backgrounds, indicating horizontal dissemination, primarily by phages. Using random mutagenesis, we recreated variation observed in databases and extended the sequence diversity of the group considerably. Characterisation of the generated variants revealed that the anti-CRISPR is tolerant to the introduction of diverse mutations, but also identified residues essential for its function predicted to impact the stability of the protein. Our analysis of the genome data publicly available and variants generated in this study, suggest that the dominant AcrIF7 type corresponds to both the optimal and minimal functional unit, possibly explaining the conservation versus mutational tolerance paradox. Our study provides a blueprint to investigate molecular evolution in other anti-CRISPR families.

INTRODUCTION

Bacteria are constantly under attack by bacteriophages. As a result, they have evolved an extensive array of anti-phage defence systems such as abortive infection, restriction-modification, and CRISPR-Cas (an acronym for Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR-associated proteins) (1). CRISPR-Cas systems are a type of prokaryotic adaptive immune system in which specific sequences targeting foreign nucleic acids are integrated into the bacterial chromosome and provide immunisation against exogenic mobile elements (2–5). These systems, widespread amongst bacteria (6), are based on the recognition of the invader nucleic acids, followed by its degradation by an endonuclease. In the opportunistic pathogen *Pseudomonas aeruginosa*, three different subtypes of the CRISPR-Cas system have been described to date: I-C, I-E and I-F, with the latter being the most abundant (7, 8).

In response to the infection barriers deployed by their hosts, phages have also developed a series of mechanisms to evade bacterial defence systems, leading to a dynamic evolutionary arms race (9). One strategy that phages have evolved to circumvent CRISPR-Cas immunity is the use of proteins that can block the system, known as Anti-CRISPRs (Acr). These proteins, first described in *P. aeruginosa* phages evading the type I-F and I-E systems (10, 11), have been identified in mobile genetic elements such as phages and plasmids, but are also encoded in bacterial genomes. In fact, it has been proposed that Acr proteins are present in more than 30% of *P. aeruginosa* strains that carry CRISPR-Cas systems (8).

Since their discovery, anti-CRISPR research has centred mainly on identifying new anti-CRISPR genes in diverse bacterial species (12, 13). Around 90 anti-CRISPR families that evade different types of CRISPR-Cas systems have been reported (14) (http://cefg.uestc.cn/anti-CRISPRdb/). These families rarely share sequence similarity and seem to possess distinctive molecular mechanisms of action. As the number of reported anti-CRISPRs keeps rising, efforts have been put into compiling and organizing anti-CRISPR sequence and metadata information in the form of a resource database (12). In its first version, anti-CRISPRdb contained 432 entries, including both experimentally validated and computationally predicted anti-CRISPRs. Currently, this database holds information for more than 3,600 anti-CRISPR proteins corresponding to at least 85 Acr sub-types (http://cefg.uestc.cn/anti-CRISPRdb/).

In comparison to the comprehensive portrait depicted for the diversity of Anti-CRISPR families thus far, the mechanisms of action of these powerful molecules have been studied to a much lesser extent. Most of the Acr mechanisms characterised to date involve blockage of different steps in the CRISPR-Cas restriction process: DNA binding, cleavage, crRNA loading, or formation of the effector-complex (15–17). For example, AcrIF1 binds Csy3 blocking the recognition of the invading DNA (15), AcrIF2 and AcrIF7 interact with the Csy1-Csy2 complex impeding DNA binding (15, 18), whereas AcrIE1 and AcrIF3 associate with the Cas3 nuclease inhibiting DNA cleavage (16). Additionally, a few structural studies of anti-CRISPR proteins have described the inter-protein interactions occurring with different components of the CRISPR-Cas system. Some of these reports have addressed the identification of important residues in the protein, typically by changing polar amino acids to non-polar ones (such as alanine) by site-directed mutagenesis (18, 19).

Along with the molecular mechanisms, the evolution of anti-CRISPR families has remained largely unexplored. Little is known about the distribution and sequence diversity of anti-CRISPRs belonging to the same family. Yet, these attributes are key to understanding how anti-CRISPRs of a certain type are acquired, what their host range is, to what extent their sequences have changed and whether such changes may impact the protein function.

Here, we use AcrIF7, an anti-CRISPR sub-type scarcely characterised, as a model to study the molecular evolution of an anti-CRISPR family. We explored the AcrIF7 diversity and distribution by searching for homologs in publicly available bacteria, plasmid and phage genomes and comparing their sequences, identifying different sub-clusters and prevalent variants. We show that AcrIF7 homologs display high sequence similarity and can be identified in diverse genomic regions, mainly in *P. aeruginosa*, where

phages appear to be major dissemination agents. Using directed evolution, we generated observed and novel AcrIF7 variants to investigate the impact of sequence variation in the anti-CRISPR function and identifying important protein residues. Our variants' characterisation identified key residues in the protein, but also distinguished regions contributing to the mutational robustness of the anti-CRISPR, thus revealing new traits of AcrIF7. Altogether, our findings imply that the dominant AcrIF7 variant represents both the optimal and minimal functional unit of the group. Our study uncovers features of AcrIF7 such as optimal sequence and length that can be used in favour of its development for biotechnology applications. Furthermore, it serves as a blueprint to investigate the molecular evolution of other anti-CRISPR families.

METHODS

Bacterial strains, phages, and culture conditions

PA14 wild-type strain, a mutant lacking the CRISPR-Cas system (PA14 ΔCRISPR loci Δcas genes, also referred to as PA14ΔCR), and the phage JBD18 were kindly provided by Professor Alan R. Davidson (20). Phage H70 harbouring *acrIF7* gene (named *g2* in the annotation of the phage genome) was isolated from a clinical *P. aeruginosa* strain (21) and belonged to Dr Gabriel Guarneros' phage collection, along with phages Ps45 and H68. Overnight cultures were grown routinely in LB (Lennox) broth with shaking at 37°C unless otherwise indicated.

Phage propagation and purification

Phages H70 (G2 carrier) and JBD18 (CRISPR-sensitive) were propagated and purified following the protocol previously reported by Cazares *et al.* (21). Briefly, the phages were propagated using the standard soft agar overlay method, followed by concentration with PEG and purification by CsCl gradient centrifugation.

Analysis of AcrIF7 sequences

The amino acid sequence of the anti-CRISPR G2 from phage H70 was first compared to all sequences available in anti-CRISPRdb (12) in August 2020 using BLASTp (22). Only proteins of the AcrIF7 family matched G2. The 68 sequences of proteins of the AcrIF7 family available in the database were then clustered using CD-HIT with an identity threshold of 100%, word size of 5 and length difference cutoff of 0 to remove identical sequences. The search of sequences homologous to G2 was extended to 50,457 proteins encoded by 574 Pseudomonas phage genomes available in GenBank, and 32,262,482 proteins from 5,279 *P. aeruginosa* genomes deposited in the GenBank RefSeq database in August 2020. A maximum e-value of 1e-03 was considered to identify homologs in all the BLASTp searches.

The amino acid sequences of G2, the 119 homologs identified in *Pseudomonas* phages and *P. aeruginosa* genomes, and the 25 non-redundant AcrIF7 sequences retrieved from anti-CRISPRdb, were aligned using the PRALINE algorithm (23) with default settings. A neighbour-joining tree was inferred from the multiple sequence alignment using the BioNJ method integrated into Seaview v4.6 (24) with 1,000 bootstrap replicates, observed distance, and including gap sites. The resulting tree was visualized with iTOL v5.7(25). An alignment of non-redundant sequences selected from the tree (deduplicated with CD-HIT and aligned with PRALINE as described above) was visualized and edited with Jalview v2.11.1.3 (26). Protein sequences representative of each sub-cluster identified in the neighbour-joining tree were scanned against all member databases in InterPro using InterProScan v5.50-84.0 (27) with default settings. The search for homologs in non-*P. aeruginosa* genomes was performed at the nucleotide level in the BLASTn suite online (22) with default search parameters, excluding *P. aeruginosa* (taxid:287) in the organisms list, and using a representative of each AcrIF7 sub-cluster as a query. AcrIF7 homologs in plasmids were searched with BLASTn against the pATLAS database (28).

Nucleotide sequences of *g2* homologs from the sub-clusters sc1 and sc2 identified in the tree were extracted from the corresponding genomic fasta files and aligned at the protein level with Seaview v4.6

(24) using the PRALINE alignment as a template. The resulting multiple sequence alignment at the nucleotide level was used in the positive selection analysis (see below). Additionally, a sequence stretch including the *acrIF7* CDS region plus 5 kb of the upstream and downstream sequence was extracted from phage genomes and *P. aeruginosa* genome sequences in the "complete" and "chromosome" GenBank assembly categories identified as carrying a G2 homolog during the initial BLASTp searches. The extracted sequences were compared with BLASTn (22) and the pairwise comparisons visualized with the genoPlotR package v0.8.11 (29).

Multilocus sequence typing (MLST) analysis

The MLST profiles of the bacterial genomes carrying AcrIF7 were identified from the pubMLST *P. aeruginosa* scheme (http://pubmlst.org/paeruginosa/) (30) using the mlst tool v.2.8 (https://github.com/tseemann/mlst).

Positive selection analysis

Positive selection analysis was performed on 131 AcrIF7 variants, including G2, the previously reported AcrIF7 (accession number ACD38920.1), and the rest of AcrIF7 variants from sub-clusters sc1 and sc2, using codeml from the PAML package v4.9 (31). The nucleotide alignment was trimmed in the 5' end to the length of G2 to have an alignment of the core codons (Supplementary data - AcrIF7 alignment). The tree used for the analysis was obtained from IQTree v1.3.11 (32), with the model K2+G4. We fit the data to the site models M1a (NearlyNeutral) and M2a (PositiveSelection). We then performed a likelihood ratio $\chi 2$ test of both models and determined the p-value. Finally, Empirical Bayes (EB) (31) was used to calculate the posterior probabilities for site classes and identify dN/dS values for each codon.

Random mutagenesis

Error-prone PCR (33) was used to introduce mutations in the sequence. Three different conditions were used (Supplementary Table 2), which differ in the concentration of MgCl2, MnCl2, and the number of extension cycles. PCR products were run in a 1% agarose gel to confirm the amplification of *g2* under the mutagenic conditions.

Cloning

The PCR products were purified using Sap-Exo kit, Jena Bioscience, and subsequently cloned into a modified version of pUCP24 plasmid (Supplementary Figure S1) in the restriction sites EcoRI and HindIII. Chemically competent DH5α *E. coli* cells were prepared and transformed with the ligation reaction following the protocol previously reported by Green & Rogers (34).

Plasmids pool preparation and electroporation in *P. aeruginosa*

Around 900 *E. coli* colonies growing in gentamicin (selective marker of the pUCP24 plasmid) were picked and grown overnight in LB-Gm (15 µg/ml). Cultures were mixed in pools of 10 candidates and plasmids were extracted using Wizard® Plus SV Minipreps DNA Purification System, Promega, to have pools of plasmids with diverse variants of *g2*, although some empty vectors were also present in the mix. To select pools with plasmids harbouring *g2*, a PCR with oligos flanking the multiple cloning site (Supplementary Table 1 - MCS pUCP24 forward and MCS pUCP24 reverse) was performed for each pool. Pools containing *g2* were selected and electroporated into *P. aeruginosa* PA14 following the protocol described by Choi et al. (35).

Selection of *P. aeruginosa* candidates carrying *g2* by colony blot

We established a colony blot protocol for detecting the presence of genes in *P. aeruginosa* in scale using a radioactive probe (Supplementary Figure S2). *P. aeruginosa* colonies carrying *g2* (not empty vectors) were selected by colony blot to make the screening more efficient than by PCR. A hundred candidates were streaked on two LB-Gm (50 µg/ml) plates (a master plate and a replica plate), along with a negative and a positive controls (colony with pUCP24-L3 empty and pUCP24-L3(*g2*), respectively), and incubated

overnight at 37°C. Colonies were then transferred from the replica plate to a nylon membrane. Membranes were placed onto filter paper damped with solution I (0.5M NaOH, 1.5M NaCl) for 10 min, and then placed onto another filter paper damped with solution II (1M Tris-HCl pH 7.2) for 2 min to neutralise the reaction. Finally, the membranes were placed on the top of filter paper moistened with solution III (0.5M Tris-HCl, 1.5M NaCl) for 5 min, and exposed to UV light for 5 min in a crosslinker to fix the DNA to the membrane. Once the DNA was fixed, the membranes were introduced in hybridisation tubes and 10 ml of a pre-hybridisation solution (1% SDS, 1M NaCl) were added and incubated at 42°C for 2 hours. During this incubation period, an oligonucleotide complementary to *g2* (G2 exp forward, Supplementary Table 1) was labelled with radioactivity (Phosphorus-32) following the protocol reported by Novogrodsky *et al.* (36). The probe was then added to the membranes and incubated at 50°C overnight. The membranes were subsequently washed with 10 ml of 2X SSC for 3 min at room temperature, followed by 2 washes with 2X SSC containing 1% SDS for 5 minutes each. The membranes were then washed for 30 min with 10 ml of 1X SSC solution, and finally with 10 ml of 0.5X SSC for 15 min. The membranes were allowed to dry before placing them onto an x-ray film in a film cassette. The film was developed, and dark spots on the film produced by the radioactive probe indicated the presence of *g2* in the colony (Supplementary Figure S2). Around 300 candidates were found to be positive for *g2.*

Phage infection assay

A hundred microliters of overnight cultures of *Pseudomonas aeruginosa* colonies carrying *g2* were mixed with 3.5 mL of TΦ top agar (1% peptone, 0.5% NaCl, 0.7% agar, 10mM MgSO4) and poured over TΦ plates (1% peptone, 0.5% NaCl, 1.5% agar) containing 50 µg/ml of gentamicin. Serial dilutions of JBD18 phage stock were spotted onto the lawns and the plates were incubated overnight at 37°C. The efficiency of plating (EOP) was calculated as the titre of the JBD18 phage in PA14 carrying the variant of G2 divided by the titre in PA14 harbouring the wild-type version of G2. Each infection assay was performed in triplicate. Welch's t-test was used and p-values were corrected for multiple comparisons using the Bonferroni correction. Adjusted p-values lower than 0.05 were considered statistically significant (Supplementary Table 3).

Sequencing and analysis of variants

Colony PCR of the colonies of *P. aeruginosa* carrying the variants of *g2* was performed using the oligos MCS pUCP24 forward and MCS pUCP24 reverse (Supplementary Table 1). The PCR products were cleaned using Sap-Exo kit, Jena Bioscience, according to the manufacturer's specifications and the variants were sequenced using the BigDye™ Terminator v1.1 Cycle Sequencing Kit, ThermoFisher Scientific. The quality of electropherograms was assessed (Supplementary Data - Electropherograms) and the mutations in G2 were determined by aligning the sequences to the wildtype version (Figure 6).

Site-directed mutagenesis

Site-directed mutagenesis of *g2* was done using the Q5® Site-Directed Mutagenesis Kit, New England Biolabs, according to the manufacturer's specifications. To generate multiple changes of amino acids in a single position, the primers were designed with one or two random nucleotides in the codon of the target residue. Oligo design was performed using the NEBaseChanger tool (http://nebasechanger.neb.com/) and the sequences can be found in supplementary table 1 (primers V40, D29 and Y32). Mutagenesis reactions were transformed into *E. coli* DH5α as previously described and subsequentially electroporated into *P. aeruginosa* PA14 to be phenotypically characterised.

Protein modelling and structural analysis

G2 wild-type and mutant carrying a deletion of the C-terminus (L55*) were modelled using the software I-TASSER (37). Analysis of the models and superpositions was done using open-source Pymol (Schrodinger, LLC. 2010. The PyMOL Molecular Graphics System, Version 2.4.0).

RESULTS

Determination of anti-CRISPR function of G2 belonging to the AcrIF7 family

Genomic analysis of phage H70 isolated from the *P. aeruginosa* clinical strain HIM5 (21) revealed the presence of proto-spacers in the ORFs 14 and 28 matching regions of the *P. aeruginosa* PA14 CRISPR loci (Figure 1a). However, infection assays showed that phage H70 could infect the strain PA14 (Figure 1b). Analysis of the H70 accessory genome identified an anti-CRISPR locus in the region of genomic plasticity (RGP) G composed of the genes *g2* and *g9* (21), which are homologous to *acrIF7* and *aca1*, respectively (38). AcrIF7 was first reported as an 83 aa protein (GenBank accession ACD38920.1) with anti-CRISPR activity against the CRISPR-Cas system I-F (12, 38). In comparison, G2 of phage H70 (Accession YP_009152337.1) is 67 aa long, lacking 16 amino acids in the N-terminus regarding ACD38920.1.

To test the functionality of this shorter version of the anti-CRISPR, we cloned *g2* in the plasmid pUCP24-L3 (Supplementary Figure S1) and assessed the protection of CRISPR-sensitive phages against the CRISPR-Cas system. The infection assays were performed in the strains PA14 wild-type (WT), a PA14 mutant lacking the CRISPR loci and *cas* genes (PA14 ΔCR), PA14 carrying the plasmid with g2 cloned (PA14-pUCP24-L3(g2)), and PA14 transformed with the empty vector (PA14-pUCP24-L3). Phage H70 was able to infect all the strains with similar efficiency, regardless of the presence of the CRISPR-Cas system (Figure 1B). In contrast, CRISPR-sensitive phages Ps45, H68, and JBD18, only produced lytic plaques in the mutant PA14 ΔCR or in the strain PA14 WT transformed with *g2*, thus demonstrating that this shorter version of AcrIF7 is a functional anti-CRISPR (Figure 1B).

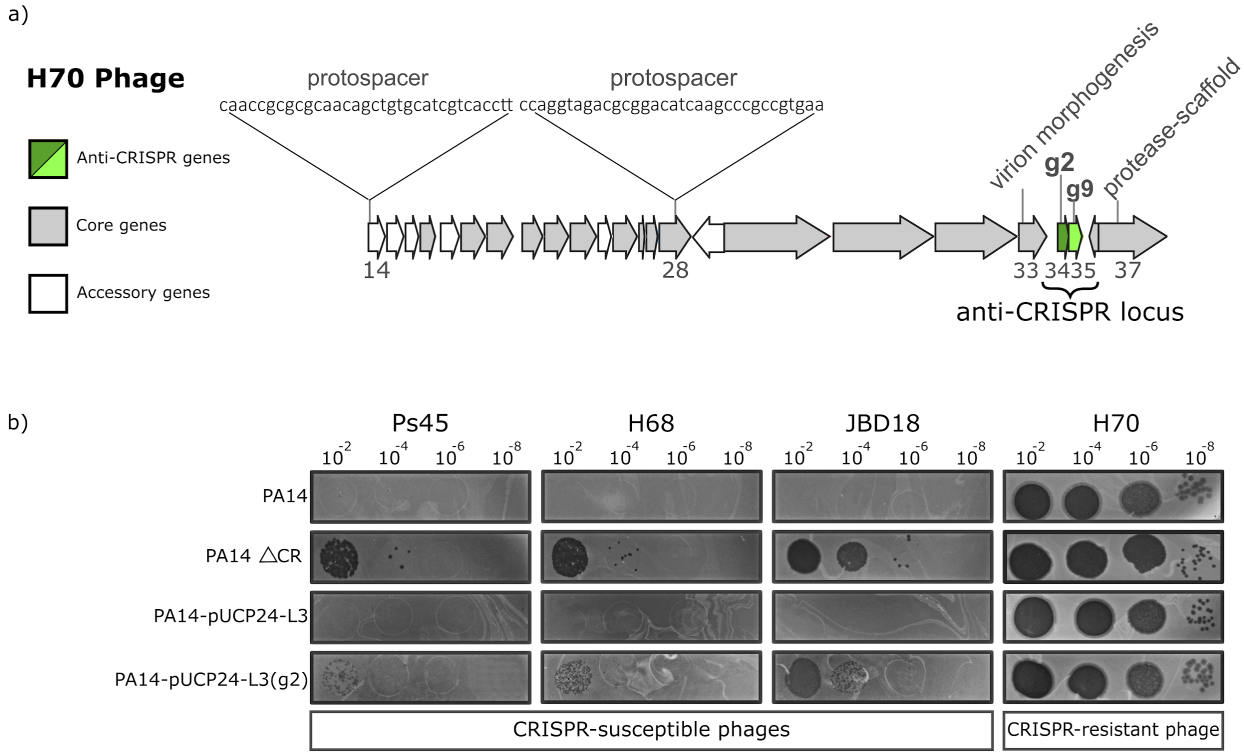

Figure 1. Location of the anti-CRISPR gene *g2* in the genome of phage H70 and inhibition of the CRISPR-Cas system I-F. A) The map represents a region of the H70 phage genome (ORFs 14 to 37 shown as arrows). The grey arrows correspond to core genes conserved in the phage group D3112virus, whereas the white arrows represent accessory ORFs (21). The anti-CRISPR locus, encoding the anti-CRISPR gene (*g2*) and a putative DNA binding gene (*g9*), are shown in green. B) Serial dilutions of different CRISPR-sensitive phages (indicated above the figure) were spotted on bacterial lawns of the PA14, PA14 ΔCRISPR-cas (PA14 ΔCR), PA14-pUCP24-L3, and PA14-pUCP24-L3(*g2*) strains. Phage infection (shown as plaques) denotes a lack of CRISPR-Cas defence due to

either the absence of the CRISPR-Cas system (PA14 ΔCR) or anti-CRISPR activity (PA14-pUCP24-L3(*g2*)). Note that the titre of each phage stocks was different, and therefore not comparable between phages.

Diversity and distribution of the anti-CRISPR family AcrIF7

Since the major difference between G2 and the first reported AcrIF7 is the additional amino acids in the N-terminus of ACD38920.1, we sought to investigate the diversity of this anti-CRISPR family. A comparative search against the ~3,600 sequences available in anti-CRISPRdb (12) only identified proteins of the AcrIF7 family as homologous to G2. The 68 homologs identified in the database, corresponding to 25 different sequences (Figure 3), included a pair of chimeric proteins with homology to anti-CRISPR of the IE4 and IF7 families. Only two AcrIF7 proteins in anti-CRISPRdb are reported as verified: anti_CRISPR0022, corresponding to ACD38920.1, the first member of the family identified in a *P. aeruginosa* genome, and anti_CRISPR0486, an AcrIE4-F7 hybrid encoded in a *P. citronellolis* genome. Members of the AcrIF7 family reported in the database are mostly associated with *P. aeruginosa* genomes, but also reported in *P. citronellolis* and different *Janthinobacterium* species (Figure 2). Yet, the AcrIF7 reported in *P. citronellolis* corresponds to an AcrIE4-F7 hybrid and the sequences from *Janthinobacterium sp.* are homologous to a protein (accession: WP_034755374.1) reported as non-functional as anti-CRISPR against the systems IF and IE of *P. aeruginosa* (38, 39).

To further explore the diversity and distribution of the AcrIF7 family, we expanded our homology search to all proteins encoded in *Pseudomonas* phage genomes deposited in GenBank and *P. aeruginosa* genomes available in RefSeq. One hundred nineteen homologs were identified, predominantly in *P. aeruginosa* genomes. Multilocus sequence typing (MLST) analysis of the bacterial sequences identified 44 different types in 94 genomes, with the remaining 23 sequences missing one or two alleles, thus indicating diversity in the *P. aeruginosa* isolates carrying anti-CRISPR of the AcrIF7 family (Supplementary Table 4). Comparison of the newly-identified homologs with the non-redundant AcrIF7 protein sequences from anti-CRISPRdb and that from phage H70 uncovered five sub-clusters within the family (Figure 2). Three sub-clusters (sc1, sc2 and sc4) were detected in *P. aeruginosa* and *Pseudomonas phage* genomes, with sc1 representing the dominant type. A BLAST search at nucleotide level against non-*P. aeruginosa* bacterial and phage genomes in GenBank using a representative of each subcluster only identified a match of sc3 in *P. citronellolis* (accession: CP015878.1) and confirmed that sc5 is associated with *Janthinobacterium* species. No matches were detected in plasmid sequences reported in pATLAS (28).

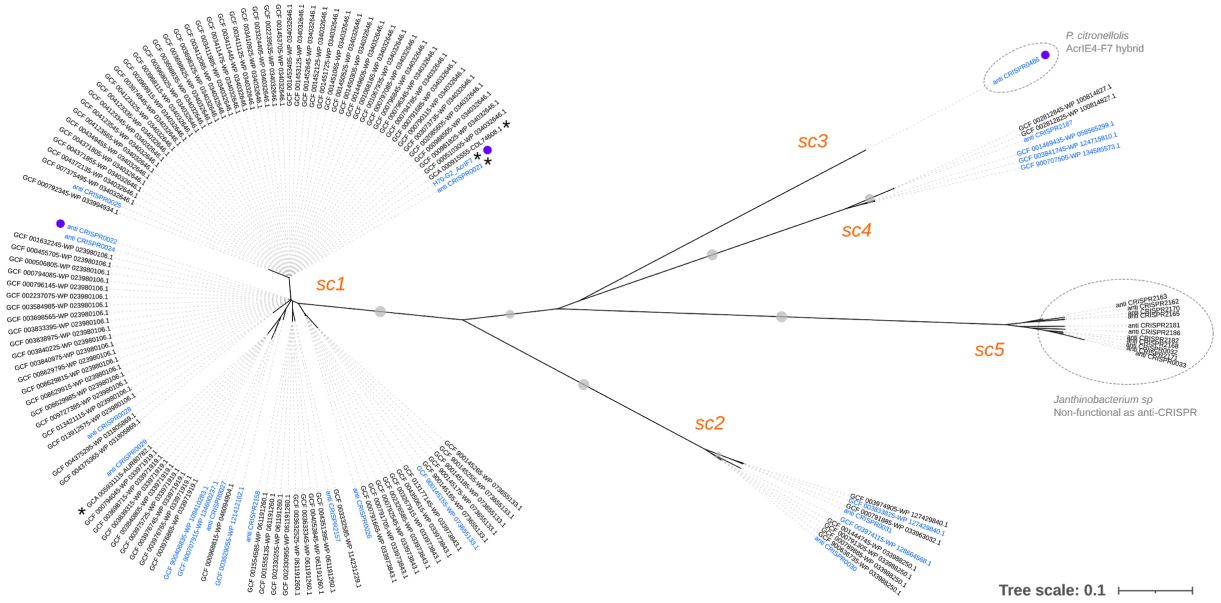

Figure 2. Diversity of members of the anti-CRISPR family AcrIF7. Neighbour-joining unrooted tree displaying the patterns of sequence similarity among protein sequences homologous to G2 of phage H70. Homologous sequences were identified through BLASTp searches against anti-CRISPRdb (12) and proteomes of *Pseudomonas* phages and *P. aeruginosa* genomes from GenBank (see Methods). The amino acid sequences of the 145 homologs presented in the tree were aligned with PRALINE (23). The tree was inferred from the resulting alignment with Seaview v4.6 (24) (BioNJ method). Grey dots on tree branches represent bootstrap support values >80 calculated from 1000 replicates. Sub-clusters (sc) identified in the tree are indicated in orange. The 25 non-redundant sequences from anti-CRISPRdb included in the tree are labelled with their corresponding identifier in the database ("anti_CRISPR" prefix). The remaining sequence labels indicate the GenBank assembly identifier and protein accession number separated by a hyphen ("-"), except for the sequence corresponding to G2 of phage H70. Asterisks mark sequences identified in phage genomes, whereas purple dots pinpoint sequences that have been experimentally verified as an anti-CRISPR. Labels in blue denote non-redundant sequences within their corresponding sub-cluster (excluding those in sc5); hence, they represent the diversity of protein sequences in the tree. Dotted line circles indicate sequences identified in non-*P. aeruginosa* genomes. Notes on the hybrid nature of the sequence in sc3 and the lack of identifiable anti-CRISPR activity against the systems IF and IE of *P. aeruginosa* in an homolog (accession: WP_034755374.1) of sc5 correspond to references (39) and (38).

Excluding sc5, for which no anti-CRISPR activity was detected in previous experimental characterisation (38), protein sequence similarity ranged from 62 to 81% between sub-clusters (Supplementary Table 5). Notably, AcrIF7 members of the same sub-cluster display limited sequence variation, with similarity values ranging from 98 to 100%, corresponding to 5-14 mutations (Figure 3). In the dominant type sc1, containing 117 members and including G2 of phage H70, only 15 different sequences and 14 mutations were distinguished (Figure 3). Sequences within this sub-cluster varied from 59 to 87 amino acids long, with 67 aa representing the most predominant length. Remarkably, the prevalence of variants within the dominant AcrIF7 type sc1 also differed considerably; variants represented by G2 and anti_CRISPR0024, separated by a single mutation, corresponded to 43 and 17% of the members in the sub-cluster, respectively (Figure 3).

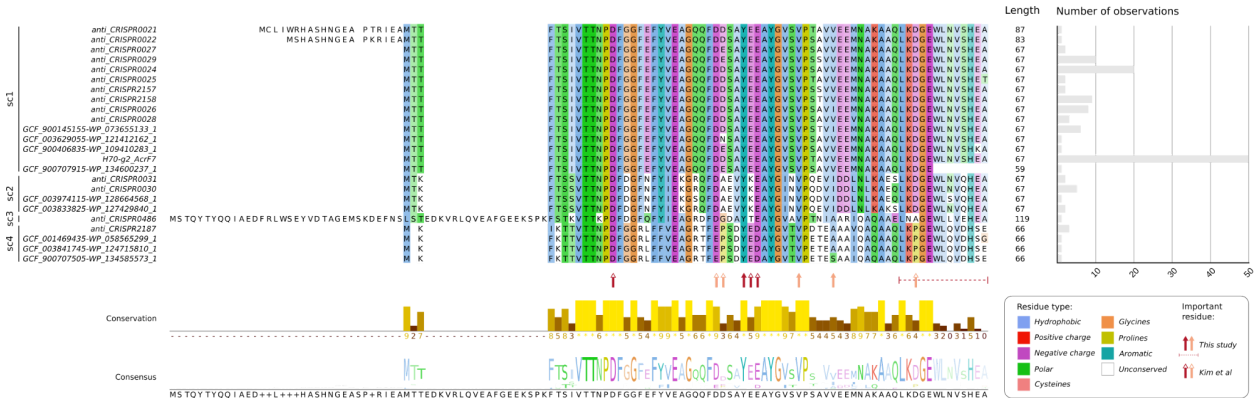

Figure 3. Alignment of non-redundant protein sequences of the AcrIF7 family. The 24 proteins selected as representative of the sequence diversity observed amongst *acrIF7* homologs (see sequence labels in blue in Figure 2), were aligned with PRALINE (23). The resulting alignment was visualised with Jalview v2.11.1.4 (26). Identifiers of the homologous variants, shown on the left side of the alignment, correspond to those described in Figure 2. The sub-cluster to which the variant belongs is indicated next to its identifier. The length of each variant sequence is displayed on the right side of the alignment, next to the bar plot illustrating the number of observations of the different variants amongst the genomes where a G2 homolog was identified (see Figure 2). Residues in the alignment are colour coded based on their level of conservation in a given position and the residue type they belong to according to the ClustalX shading scheme, indicated at the bottom-right of the figure. The conservation level and consensus sequence of the alignment are represented with a bar plot and sequence logo at the bottom of the figure, respectively. Residues identified in this study as important for the G2 anti-CRISPR activity are pinpointed with solid arrows or a dotted line below the alignment. The dotted line indicates that the lack of the underscored residues in G2 nullifies the anti-CRISPR activity of the protein (see Figure 6). Residues important for the AcrIF7-CRISPR-Cas interaction as reported by Kim et al. (18) are identified with open arrows. Red arrows denote residues on which mutations drive the loss of the AcrIF7 function or interaction whereas orange arrows indicate residues on which mutations have a partial effect.

Since *g2* is encoded in a phage genome, we investigated whether sequence conservation observed amongst homologs of the same sub-cluster was extended to the genomes carrying the anti-CRISPR gene. We extracted and compared regions flaking anti-CRISPR of the AcrIF7 family identified in phage and complete *P. aeruginosa* genomes. The comparison revealed that anti-CRISPRs AcrIF7 can be associated with different genomic backgrounds (Figure 4). Members of the sub-cluster sc1 were detected in three distinct phage types and two divergent bacterial genome regions, indicative of anti-CRISPR acquisition via horizontal gene transfer. Still, AcrIF7 members of the sub-cluster sc1 were largely linked to transposable phages of the group D3112-like (Figure 4). In terms of neighbour genes, no other anti-CRISPRs were detected next to the *acrIF7* gene in the regions analysed, but *aca1* was typically located immediately downstream (Figure 4). Surprisingly, the comparative analysis also uncovered the presence of more than one copy of this anti-CRISPR gene in the genome of three *P. aeruginosa* strains: RIVM-EMC2982 (GCF_002085605), Carb01 63 (GCF_000981825) and NCTC11445 (GCF_900636735). In these cases, the *acrIF7* gene was located in different genome positions, and it was frequently surrounded by homologous phage genes. Despite the existence of sequence variation between the homologous regions detected in the comparative analysis, the anti-CRISPR coding region commonly bore no mutations or displayed nucleotide sequence conservation above the average (Figure 4). For example, the 204 bp coding region of H70 *g2* was identical in 11 genomes, including in the unrelated regions of *P. aeruginosa* SCV20265 and phage vB_Paep_Tr60_Ab31. Such levels of sequence similarity hampered the discrimination of conserved positions in the AcrIF7 protein sequence that could be important for the anti-CRISPR function. Moreover, screening of the diversity of AcrIF7 sequences for the presence of known protein signatures or classification into established protein families yielded no results; hence, no functional domains or motifs, nor links to functionally characterised protein sequences were identified, underscoring the novelty of the AcrIF7 family.

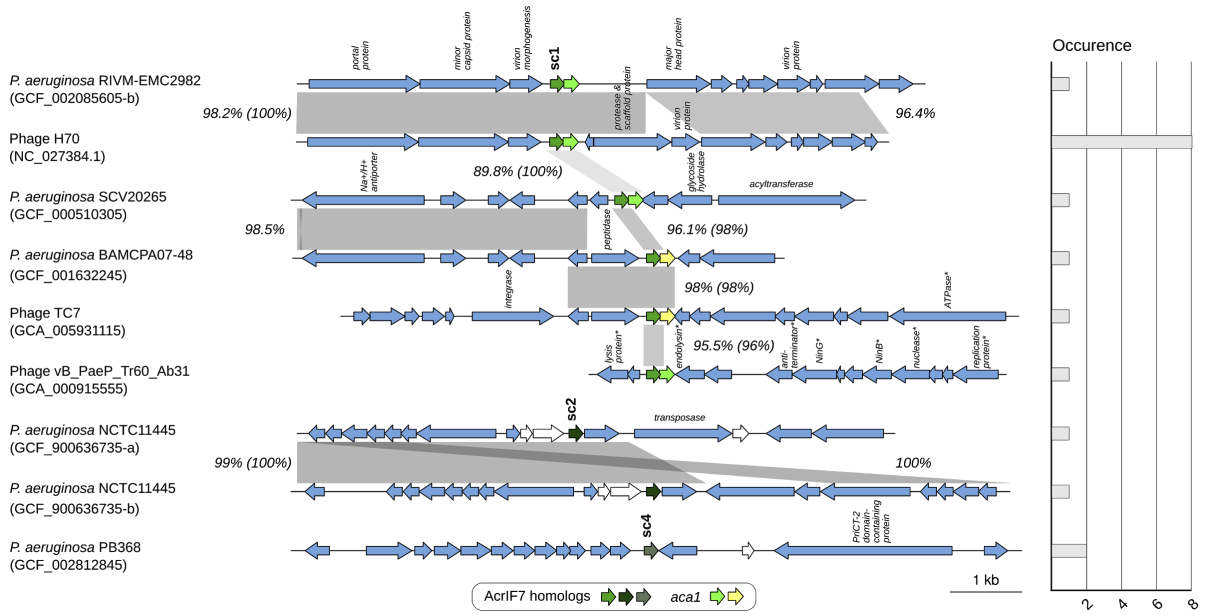

Figure 4. Comparative analysis of genome regions harbouring AcrIF7 homologs. The figure shows pairwise comparisons at the nucleotide level of regions containing an AcrIF7 homolog in complete *P. aeruginosa* and phage genomes. Regions containing an AcrIF7 homolog gene plus 5 kb of upstream and downstream sequence (where available), were extracted and compared all-vs-all with BLASTn. Regions were then clustered based on sequence similarity to determine their frequency of occurrence among the set of analysed genomes. We defined regions as belonging to the same type when sharing >95% overall sequence coverage and identity. The figure summarises the comparison of 17 regions from 13 genomes, displaying the similarity between representatives of the different region types identified. Regions are paired with their closest match and their occurrence is illustrated with a bar plot on the right side of the figure. The organism name and GenBank accession number (in parenthesis) of the genomes from which the regions were extracted are indicated next to their corresponding gene maps, on the left side of the figure. Instances where more than one AcrIF7 homolog was detected in the same genome are distinguished with a suffix letter added to the GenBank accession number. The AcrIF7 homolog genes, and *aca1*, are colour coded as indicated in the figure. The sub-cluster type of the different AcrIF7 homolog genes (see Figure 2), is shown above the corresponding gene arrow. When available, functions assigned to ORF products, as indicated in the genbank file, are displayed above the corresponding arrow. Light yellow arrows denote ORFs encoding homologs of Aca1 overlooked in the original annotations. White arrows indicate overlooked ORFs with unknown function. The percentage of sequence identity detected between homologous regions, depicted as grey connecting blocks, is indicated next to the corresponding block. For homologous regions containing an AcrIF7 homolog gene, the percentage of identity between the gene sequences is additionally indicated in parenthesis.

Positive selection analysis

Most AcrIF7 homologous proteins identified in databases displayed high sequence conservation (Figures 2 and 3), yet, some positions exhibiting variation were observed in the alignment of non-redundant protein sequences of the AcrIF7 family (Figure 3), implying distinct patterns of divergence at the codon level. We, therefore, sought to estimate the synonymous and nonsynonymous substitution rates of each residue, looking for codons potentially under positive selection that could drive protein fitness optimization. Only representatives of the sub-clusters sc1 and sc3 have been reported to be active as anti-CRISPR against the CRISPR-Cas system I-F of *P. aeruginosa* (Figure 2). We decided, however, to focus on the sequences of the sub-clusters sc1 and sc2, since the divergence of the chimeric sc3 representative at the nucleotide level could compromise the reliability of the analysis. Moreover, sc1 and sc2 comprise the most abundant sub-clusters, accounting for around 90% of all the members of the AcrIF7 family. One hundred thirty-one sequences (Supplementary Data - AcrIF7 alignment) were analyzed using the models M1a (NearlyNeutral) and M2a (PositiveSelection), which estimate the evolutionary pressures over specific codon sites in the gene (Supplementary Data - M1a/M2a). After performing a likelihood ratio $\chi 2$-test of both models, we found no significant evidence to reject the NearlyNeutral model (p-value= 0.98416),

suggesting that sequences in these sub-clusters of the AcrIF7 family are not under a positive selection. Furthermore, the Naive Empirical Bayes (NEB) results from the M1a model (Supplementary Data -M1a), which calculates the omega (dN/dS) values and their probabilities, showed that positions 29, 30, 42, 43, 45 and 64 (using g2 as reference) have an omega of around 1, predicting that they have a neutral effect on the protein despite displaying sequence variation (Figure 3).

Directed evolution of AcrIF7

Since the most abundant sub-clusters in the AcrIF7 family were highly conserved and predicted to be under nearly neutral evolution, we undertook an unbiased approach similar to directed evolution to yield insights into the protein functionality by introducing mutations in the gene and identifying those relevant for the anti-CRISPR activity. We opted to use a random mutagenesis strategy (as it is considered a powerful tool for directed evolution of proteins (40)), in which we implement error-prone conditions during the PCR to promote the misincorporation of nucleotides by the polymerase (32). Unlike common directed evolution approaches, in our method we did not aim at finding better versions of AcrIF7 as it can already completely block the CRISPR-Cas system I-F. Instead, we sought to explore the sequence landscape to find mutants with similar anti-CRISPR activity as G2, and mutants with an impared function. Three different mutagenic conditions were applied to widen the error rate spectrum (Supplementary Table 2), which ranged from 0.6 to 1.4% per position. We cloned the mutagenic PCR products into the pUCP24-L3 vector and used the resulting library to characterise functionally important amino acids in the protein and identify regions that contribute to the mutational tolerance (also known as mutational robustness) (Figure 5). After one round of directed evolution, the functionality of all the generated *g2* variants was assessed through infection assays, testing the ability of the mutant protein to block the CRISPR-Cas system and allow infection by the CRISPR-susceptible phage JBD18 (Supplementary Figure S3).

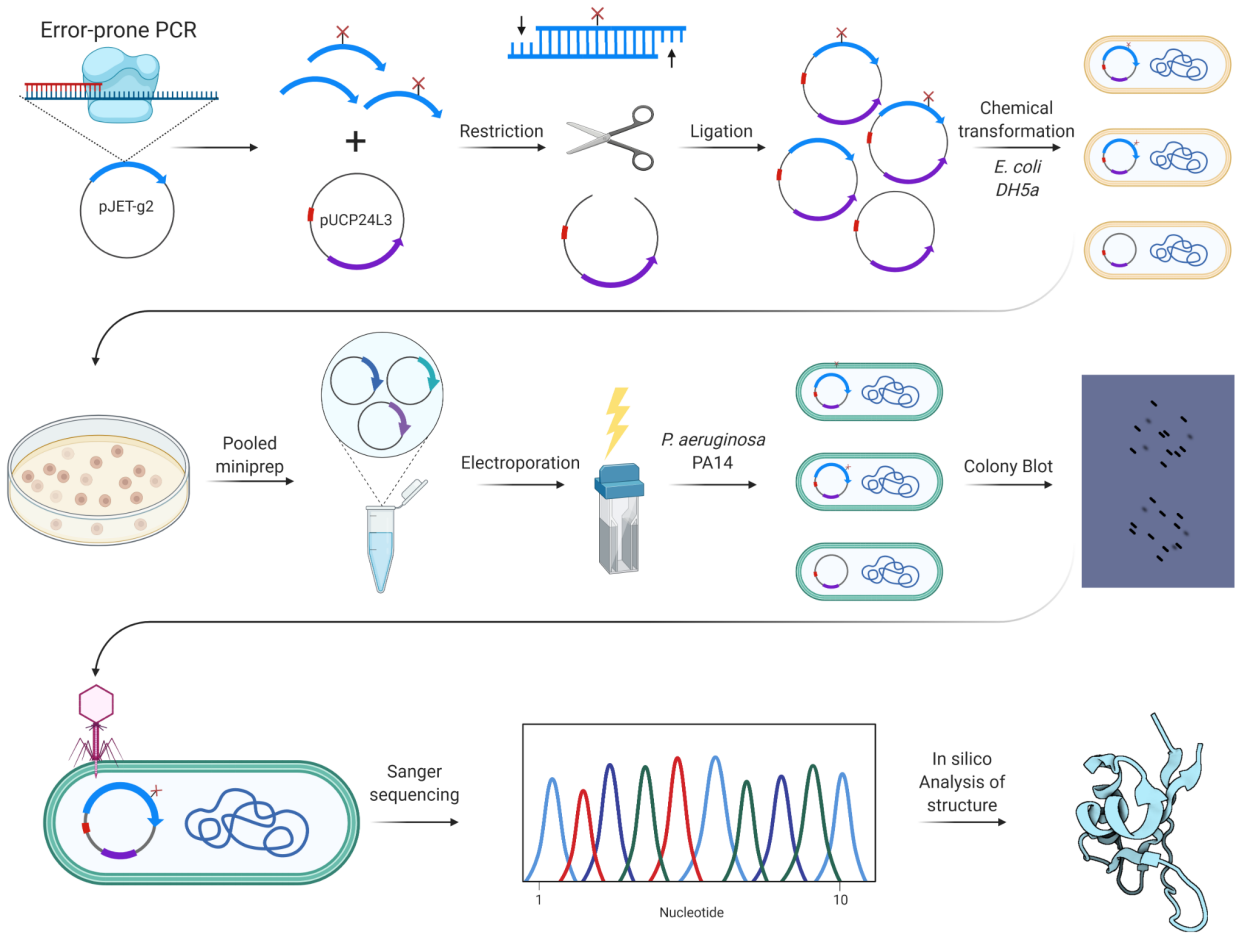

Figure 5. Strategy for cloning and identification of G2 mutants. The figure illustrates the steps followed to generate the collection of G2 random mutants presented in this study. The strategy consisted of: 1) cloning the error-prone PCR products into pUCP24-L3, 2) transformation and extraction of pools of plasmids from *E. coli*, 3) electroporation of the pools into *Pseudomonas aeruginosa* PA14, 4) assessment of the efficiency of the variant to block the CRISPR-Cas system and 5) sequencing of the mutants and analysis of the protein model. Created with BioRender.com.

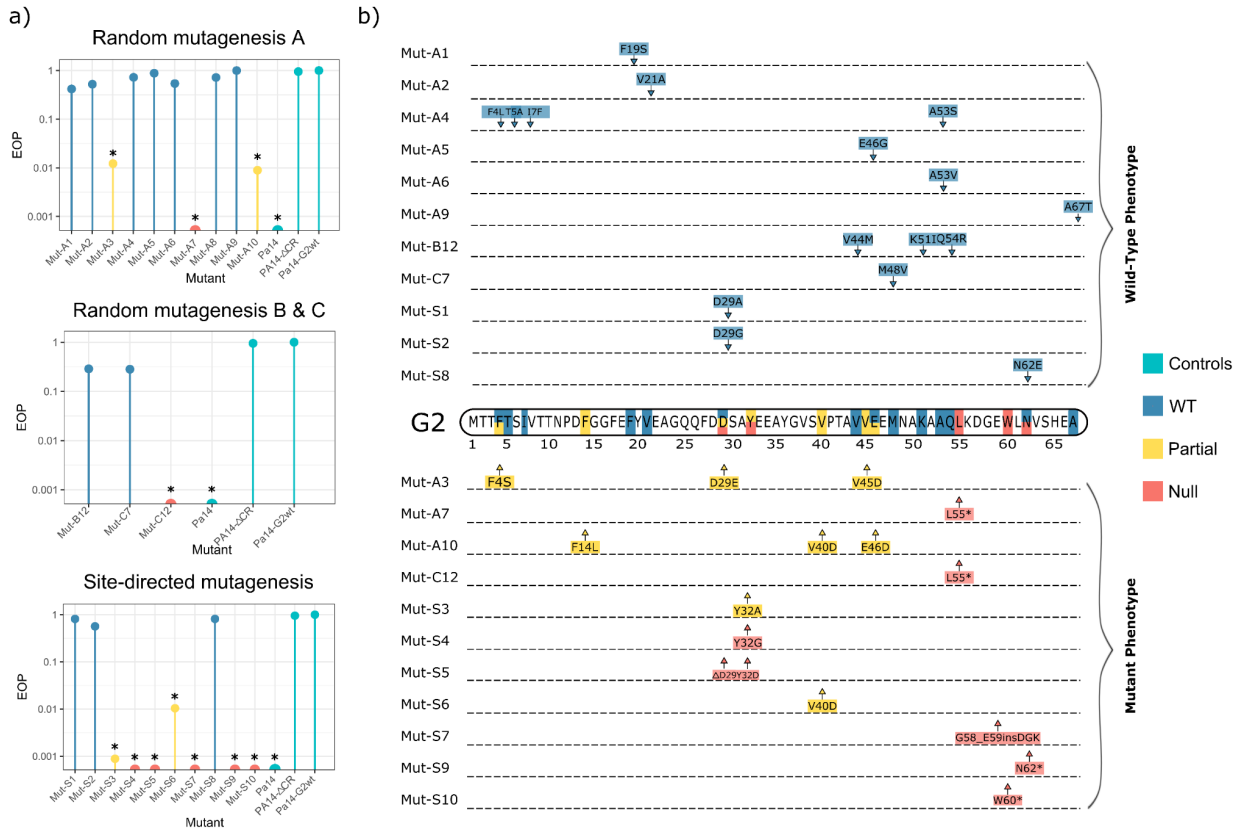

Figure 6. Efficiency of CRISPR-Cas inhibition by G2 variants and map of mutations. The efficiency of G2 mutants at inhibiting the CRISPR-Cas system I-F is shown in the left panel. The lollipop charts show the efficiency of plating (EOP) of the CRISPR-sensitive phage JBD18 on PA14 carrying different variants of G2, normalised to the titre of the same phage in PA14 featuring the WT version of G2. Asterisks denote adjusted p-values ≤0.05 (Raw data of replicates and p-values can be found in Supplementary Table 3). In the right panel, a map of the amino acid changes present in the mutants is shown. The colours represent the phenotypes of the G2 variants based on their inhibiting of the CRISPR-Cas system: wild-type (in blue), partial loss-of-function (in yellow), or null activity (in red). The changes in each mutant are shown above the dotted line next to the variant's name (e.g. Mut-A1 has a mutation in F19S, whereas Mut-A3 has mutations in F4S, D29E, and V45D). The wild-type sequence of the anti-CRISPR is displayed in the middle of the figure, with each mutated position coloured according to the phenotype of the mutant that carried changes in that position.

Most of the tested mutant candidates displayed a wild-type phenotype, i.e. JBD18 was able to infect the strain carrying the mutant with the same efficiency as when the strain was transformed with G2 wild-type (Supplementary Table 3). However, we were also able to distinguish mutants with affected anti-CRISPR activity. Only 2 random mutants displayed a partial loss-of-function (Mut-A3 and Mut-A10) whereas the other 2 featured null anti-CRISPR activity (Mut-A7 and Mut-C12) (Figure 6). Although the JDB18 efficiency of plating (EOP) decreased ~100-fold in the partially-functional mutants, the phage was able to infect the PA14 strain, indicating a residual protection effect provided by these G2 variants unable to block the CRISPR-Cas system completely. In the case of the mutants Mut-A7 and Mut-C12, no phage plaques were detected in the lawns, implying the lack of anti-CRISPR activity in the variants. No mutant candidates exhibited an improvement in the JBD18 EOP compared to G2 wild-type.

The mutant candidates with affected phenotypes, and a handful of randomly selected candidates displaying a wild-type phenotype, were sequenced to identify changes that led to the loss of function and neutral mutations (Figure 6). Mutations identified in candidates with wild-type phenotype were scattered across the G2 sequence (Figure 6). Note that from all wild-type mutant candidates detected in the

infection experiments, only those sequenced and harbouring non-synonymous mutations in *g2* are presented in Figure 6.

The two candidates displaying partial anti-CRISPR activity had 3 mutations each. Mut-A3 featured 3 punctual mutations: F4S, D29E, and V45D. In the case of the mutation in residue 29, aspartate and glutamate are both negatively charged amino acids; hence, the change between the two amino acids is less expected to have an impact on the protein function. As for the mutation F4S, a mutant with wild-type phenotype (Mut-A4) also featured a mutation in the same position, along with 2 other mutations in adjacent amino acids (T5A and I7F) (Figure 6); thus pinpointing the change in the valine 45 for an aspartic acid as the most likely driver of the reduction in anti-CRISPR activity observed in this G2 variant.

The second mutant with partial protein function, Mut-A10, acquired 3 mutations: F14L, V40D, and E46D. The mutation E46D seemed unlikely to have a strong impact on the protein function since both amino acids are chemically similar and the position was also mutated in a wild-type mutant (Mut-A5: E46G) (Figure 6). Likewise, phenylalanine and leucine, the amino acids changing in position 14, are both nonpolar despite being structurally different. In contrast, the mutation V40D seemed to be relevant, since a nonpolar amino acid was replaced by an acidic one, as observed with the mutation deemed as responsible for the partial loss of function of the mutant Mut-A3: V45D.

Both null mutants acquired mutations that resulted in the introduction of a premature stop codon in position 55 (L55*), albeit in a different manner (Figure 6). Mut-A7 featured an SNP that changed the lysine codon for a stop codon, whereas Mut-C12 had a deletion of a nucleotide that caused a frameshift leading to a stop codon gain. The premature stop codon resulted in the deletion of 13 amino acids of G2, which is already a small protein (67 amino acids long). Among the potential effects, loss of the 13 amino acids in the carboxy-terminus of the protein could alter its tertiary structure and therefore the recognition of, and interaction with, components of the CRISPR-Cas system.

Site-directed mutagenesis and protein model

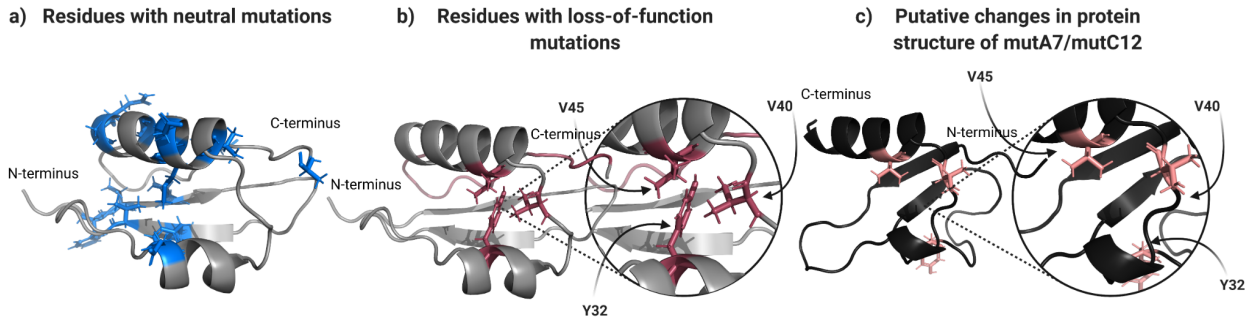

Figure 7. Model prediction of G2 and residues with a neutral or negative impact on the function. The structure of G2 was predicted using the server I-TASSER. A) Residues with mutations that displayed no impact on the anti-CRISPR function are shown in blue. B) Amino acids that were found to contribute to the protein activity are shown in red. C) Protein model prediction for the mutants mutA/mutC12 that lack 13 amino acids in the C-terminus. Amino acids in pink correspond to the same positions as those in red in panel b); the figure shows the displacement of Y32 in the structure of the mutant, while V45 and V40 remain in the same predicted position as G2 WT.

To gain further insights into the impact of the mutations affecting the G2 anti-CRISPR activity, and the relevance of the mutated residues in the protein, we predicted the three-dimensional structure model of G2 and the null mutants (Mut-A7/C12 (L55*)) using I-TASSER (37) (since no structure of AcrIF7 was available by the time we did the analysis). The computationally-generated structure of G2 featured long and short alpha helices and two anti-parallel beta sheets (Figure 7). The predicted structure of the

Mut-A7/C12 null mutants differed from that of G2, especially in the position of the short alpha helix (Figure 7), suggesting that the thirteen residues lost in the carboxy-termini of the mutants could be important to maintain the stability of the protein structure and therefore its function.

Analysis of the structure model of G2 showed that residues with identified neutral mutations, i.e. those that did not affect the anti-CRISPR activity, were dispersed in the protein structure (Figure 7, residues highlighted in blue). Residues with mutations that we predicted to be responsible for the partial loss of function observed in Mut-A3 and Mut-A10, namely V40 and V45, were located closely in the structure model (Figure 7, residues highlighted in red). Intriguingly, V40 and V45 were tightly clustered with a tyrosine located in the short alpha helix (Y32) in the interior of protein (Figure 7). Since Y32 moved far away from the cluster in the structure model of the null mutants, we hypothesised this amino acid could also be important for the anti-CRISPR function.

To test our hypotheses, we created a new series of mutants by site-directed mutagenesis. Firstly, we changed Y32 for amino acids with different chemical properties to investigate the significance of this residue in the protein function. The introduction of an alanine (mutant Mut-S3) led to a 1000-fold reduction in the anti-CRISPR activity, whereas the introduction of a glycine (mutant Mut-S4) rendered the protein inactive (Figure 6), thus indicating that tyrosine 32 is a key residue in AcrIF7. Another mutant with a change in the tyrosine 32 (Mut-S5) showed no anti-CRISPR activity as well, however, this change was preceded by the unexpected deletion of the amino acid in position 29. Next, we assessed whether the mutation V40D drove the partial loss of function previously observed in the mutant Mut-A10, which carried other two mutations. The new mutant Mut-S6, bearing V40D solely, exhibited a reduction in the EOP of phage JBD18 equal to that recorded for Mut-A10 (Figure 6), confirming that this mutation was responsible for the impairment of the anti-CRISPR activity. The partial mutant Mut-A3 also featured changes in three residues, of which D29 and V45 were not mutated in our set of wild-type mutants (see Random mutagenesis section). We found that replacing D29 with amino acids of different chemical properties (Mutants Mut-S1 and Mut_S2) did not have an impact on the anti-CRISPR activity (Figure 6), implying that the mutation V45D caused the partial loss of function in Mut-A3. Altogether, this evidence shows that Y32, V40 and V45 are key residues for AcrIF7. Together, they likely play an important role in stabilizing the structure of the protein.

Since null mutants lost 13 amino acids in the carboxy-terminus, we further investigated the importance of these residues by introducing stop codons in other positions, looking to test whether variants longer than the null mutants were still functional. From the thirteen codons, those corresponding to positions 60 and 62 were selected to turn into stop codons because they required one single mutation. Moreover, an AcrIF7 variant identified in databases had a stop codon in position 60, making it an interesting position to study. We found that introducing stop codons in positions 60 or 62 (mutants Mut-S9 and Mut-S10), leading to loss of 8 and 6 amino acids, respectively, rendered the anti-CRISPR protein inactive (Figure 6). Interestingly, changing the residue in position 62 (N62E, mutant Mut-S8) did not have an impact on the anti-CRISPR activity, implicating that loss of residues in the carboxyl terminus is the main factor driving the anti-CRISPR inactivation. Together with the analysis of the predicted structure models (Figure 7), these results suggest that 67 amino acids constitute the minimal functional AcrIF7, particularly on the carboxy-terminal portion of the anti-CRISPR, which may play a critical role in preserving the protein structure. This notion is in line with the characterization of an unexpected mutant arising in our experiments (Mut-S7), which displayed no anti-CRISPR activity as a result of a three-residue insertion in position 58 of the protein.

In summary, we introduced 30 different mutations in 21 positions scattered throughout the anti-CRISPR gene (Figure 6), and corresponding to 31.3% of the protein. Seven of the mutations generated *in vitro* were present in AcrIF7 variants identified in databases, with six of them displaying a wild-type phenotype. Notably, mutations introduced in 14 different positions, corresponding to 66.6% of the mutated residues, showed a neutral effect on the protein function, indicating that these positions contribute to the mutational tolerance of AcrIF7.

DISCUSSION

CRISPR-Cas systems are widespread in both archaea and bacteria (6), leading to the expectation that a wide variety of anti-CRISPRs exist in phage or other mobile elements to aid them circumvent this effective defence system. Consistent with this notion, numerous anti-CRISPRs have been identified recently (12), and it can only be expected that the number of reports on new anti-CRISPs will keep rising rapidly. Yet, despite more anti-CRISPRs being discovered, many questions remain regarding how these remarkable proteins work, evolve and spread. In this study, we use AcrIF7 as a model to investigate functionality, diversity, distribution and evolution within an anti-CRISPR family.

By searching for homologs of the anti-CRISPR *g2* in different databases and comparing their sequences we were able to portray a detailed picture of the diversity within the AcrIF7 family. This enabled us not only to identify different sub-clusters and characterise their levels of sequence similarity but also to distinguish prevalent types and a small number of variants representing the diversity of the group (Figures 2 and 3). Our findings indicate that AcrIF7 homologs are mainly associated with *P. aeruginosa*, possibly suggesting specialization to the CRISPR-Cas system of the species in which this anti-CRISPR class was first identified (38). One exception is the member of the sub-cluster 3, a hybrid of the anti-CRISPR families IE4 and IF7 found in *P. citronellolis*, indicative of the potential flexibility of these molecules (39). Intriguingly, we did not identify the hybrid in *P. aeruginosa* genomes, raising questions of whether this fusion arose in *P. citronellolis*, or it originated in *P. aeruginosa* but it has been negatively selected, perhaps due to the presence of other anti-CRISPRs of the class I-E in the genomes which could make the fusion redundant.

After further exploration of regions flanking AcrIF7 homologs in complete genomes, we uncovered that this anti-CRISPR type can be linked to diverse genetic backgrounds in *P. aeruginosa* (Figure 4), indicating gene mobilization. We identified AcrIF7 homologs of the sub-cluster 1 in three unrelated phages sharing sequence similarity restricted to less than 2% of their genomes: the siphophage H70 isolated from a clinical strain of *P. aeruginosa* in Mexico (21), the podophage Ab31 isolated from waste water in Ivory Coast (41) and the siphophage TC7 isolated from hospital sewage in China (see metadata in GenBank record MG707188.1). The isolation source of these phages outlines the importance of anti-CRISPRs in different environments. Moreover, the fact that the three phages are temperate, implies a key role for this phage type on the horizontal transfer of anti-CRISPRs, likely promoted by the co-existence of diverse prophages in a bacterial genome.

Although AcrIF7 homologs of the sub-cluster 1 were identified in distinct phages, they were predominantly associated with transposable phages of the type D3112-like (Figure 4). The genomes of these phages are known to contain multiple regions of genomic plasticity accommodating various accessory genes, with one region in particular harbouring diverse arrays of anti-CRISPR genes (10, 21). In fact, anti-CRISPRs were first discovered in D3112-like phages (10). Thus, it appears that this type of transposable phages represents a major reservoir of anti-CRISPR genes in *P. aeruginosa*, which could then be transferred to other mobile elements, consistent with our observation on the abundance of AcrIF7 homologs in D3112-like phages compared to other phage types. AcrIF7 homologs were also detected in divergent regions of *P. aeruginosa* genomes sharing little or no similarity with phage sequences, implying the acquisition of AcrIF7 by yet uncharacterised genetic elements. Regardless of the genetic background in which AcrIF7 homologs were identified, we observed strikingly high levels of sequence similarity among their coding regions. Unlike other anti-CRISPRs (13, 42), we did not detect AcrIF7 homologs in plasmids. Whilst this observation may hint a preferred association with phages or other types of mobile elements, it could also be explained by the limited number of plasmid records existing for *P. aeruginosa*.

The most divergent sequences in the group AcrIF7 correspond to closely related proteins encoded in *Janthinobacterium sp.* genomes, here clustered within the sub-cluster 5 (sc5; Figure 2). These sequences, however, are nearly identical to a protein that did not exhibit anti-CRISPR activity when tested against the CRISPR-Cas systems I-F and I-E of *P. aeruginosa* and I-F of *P. atrosepticum (38)*, thus suggesting that members of the sub-cluster 5 may either be active against CRISPR-Cas systems of other

species or feature an unrelated function. The fact that proteins in sc5 share 62% similarity with those of the sub-cluster 1 (Supplementary Table 5), and can therefore be easily detected as potential homologs through BLAST searches, prompts to be cautious about how we infer anti-CRISPR functions from sequence homology information. Since only representatives of the sub-clusters 1 and 3 in the group AcrIF7 have been experimentally verified as anti-CRISPRs, it remains to be seen whether proteins in the sub-clusters 2 and 4 share the same function. Together with the lack of protein signature information associated with anti-CRISPRs (e.g. protein families or conserved domains), these observations highlight the need to investigate anti-CRISPR families in detail combining both computational and experimental approaches to further our understanding of their diversity, functionality and evolution.

In vitro directed evolution, especially by random mutagenesis, coupled with a selection method is a powerful approach to characterise novel proteins for which limited information is available (43). This strategy is particularly effective at identifying not only residues involved in protein binding, but also in protein folding, or at finding mutations that enhance protein activity (33, 44, 45). This is relevant for protein families such as AcrIF7, which are highly conserved according to the sequences available in databases, making it difficult to draw inferences about the impact that certain positions have on the protein. By implementing an approach similar to directed evolution, we not only captured and experimentally characterised some of the diversity observed for AcrIF7 in databases, but generated more variation to study the functionality of a conserved anti-CRISPR. Remarkably, our mutational screening revealed important residues contributing to the anti-CRISPR functionality that could not be identified with traditional methods which are biased towards testing polar amino acids (18, 19). Additionally, this approach enabled the identification of regions that contribute to the mutational robustness of the protein.

Essential proteins are typically conserved (46). This principle may help to understand why some anti-CRISPR families feature high levels of sequence similarity. Although not all phages encode an anti-CRISPR, this function becomes essential when infecting a host with a CRISPR-Cas system. In this context, mutational robustness represents an advantageous trait that minimizes the effect of random changes on the protein function (47, 48). Despite the high levels of conservation observed in most of the members of the AcrIF7 family, our experiments identified mutations with a neutral effect on the anti-CRISPR function. These residues, contributing to the mutational tolerance of the protein, correspond to ~67% of the amino acids mutated in our study. Our findings are reminiscent of those reported for the Influenza A virus matrix protein M1, and the middle domain of the heat shock protein Hsp90 from yeast (49, 50), for which high levels of intrinsic tolerance to mutations were found despite the little sequence variation observed in nature. These three proteins play crucial roles in the survival of their carrying microorganism: AcrIF7 is essential for a successful phage infection in the presence of a CRISPR-Cas system; M1 participates in multiple stages of the viral infectious cycle; and Hsp90 is involved in protecting cells from environmental stress and growth at high temperature (49, 50). The key role of these proteins, together with a possible selective pressure *in vivo*, may explain why they are highly conserved even though they can tolerate mutations *in vitro*. Future studies focused on deep mutational scanning of anti-CRISPRs would tell us how robust they are in comparison with other proteins.

Although G2 was capable of carrying multiple mutations without having its function significantly affected (Figure 6, Supplementary Table 3), it is possible that those variants are not as stable as the wild-type version, and are therefore negatively selected; thus suggesting that the sequence of G2 represents the optimal AcrIF7 version. This notion is supported by the fact that G2 can completely block the CRISPR-Cas system I-F, i.e., a CRISPR-sensitive phage can infect a strain carrying G2 WT with the same efficiency as the mutant lacking the CRISPR-Cas system (Figure 6b). Moreover, variants with the G2 sequence are the most prevalent among the members of the family (Figure 3), possibly suggesting a selection for this version of AcrIF7.

No structure was available for AcrIF7 when we started this study. Hence, we decided to model the protein structure using I-TASSER (37) to further investigate the impact of the mutations introduced in G2. The solution of the AcrIF7 structure has been recently reported by Kim *et al*. (18) and of AcrIF7 bound to the Csy complex (51). Remarkably, the structure that we obtained from I-TASSER is highly similar to that

solved by NMR (RMSD of 3.708 Å, Supplementary Figure S4), mainly differing in some of the random coils. These results underscore the power of computational protein structure modelling, as recently highlighted by the use of Google's DeepMind to solve a protein's 3D shape (52). Here, we exploit the computationally-generated structure of AcrIF7 to guide the identification of residues important for the anti-CRISPR function. The protein model predicted for the mutants mutA7/C12 (Figure 7) was key for us to hypothesise the involvement of Y32 in the anti-CRISPR activity, which was later confirmed by site-directed mutagenesis (Figure 6).

We discovered that Tyr32 is key for the anti-CRISPR activity of G2. Studies by Kim *et al.* and Gabel *et al.* found that AcrIF7 binds the Cas8f–Cas5f complex, and that the interaction was driven by hydrogen bonds and electrostatic interactions (18, 51). Kim *et al.* found that residues E33, E34 and D13 contributed to the electrostatic interactions with Cas8f (18), whereas Gabel *et al.* discovered that D13, Y20, V21, E22, D28, E34 and S64 interact with different regions of the Csy complex (51). Based on the structure reported by Gabel *et al.* we hypothesise that Tyr32 contributes to the anti-CRISPR activity by stabilising the protein, (53). Likewise, we found that mutations in valines 40 and 45 had a significant impact on the function. Valines are extremely hydrophobic amino acids likely contributing to the hydrophobic packing of AcrIF7 (54). This highlights the importance of our *in vitro* evolution strategy to study anti-CRISPR proteins, since it allowed us to find amino acids that are not part of the protein surface and still contribute largely to its function. Our work also underscores the importance of doing functional analyses to complement structural studies, since we identified mutations in amino acids that were found in structural studies to interact with the CRISPR-Cas system (V21) (51), and yet, the changes that we found had a neutral effect on the anti-CRISPR function.

Kim *et al.* showed that changes in the position 29 (D29K), in combination with a change in the aspartic acid 28 (D28K), induce a modest increase in the equilibrium dissociation constant (K_D) of the protein in association with Cas8f-Cas5f (18). However, our study, together with the structural analysis by Gabel *et al.*, shows that D29 alone does not have an impact on the protein's function, since we found that mutations in this position (D29A and D29G) did not exhibit a reduction of the anti-CRISPR activity (Figure 6). Additionally, our evolutionary analyses showed D28 is highly conserved among the AcrIF7 variants found in databases, whereas D29 was identified as one of the least conserved positions (Figure 3) and possibly under neutral evolution (Supplementary data-M1a).

We observed that deletions in the Carboxy-terminus of G2 completely abolished the anti-CRISPR function. The structure model of the mutants with the deletion L55* (Mut-A7/C12) suggests that this region is important for the maintenance of the structure and the stability of the protein. No longer variants at the Carboxy-terminus of AcrIF7 were identified in databases. In contrast, we found that the anti-CRISPR encoded in the genome GCF_900707915 is shorter. Our results for the mutant Mut-S9 (N62*), however, suggest that this variant is no longer active. Altogether, our *in vitro* and bioinformatics analyses is consistent with the fact that the most predominant variant length is 67 amino acids (93.6% of the sc1 and nearly 65% of all sub-clusters), implying that this protein size has been evolutionarily selected.

In addition to furthering our understanding of the phage-bacteria arms race and co-evolution, the study of anti-CRISPRs can boost their application in biotechnology, similar to the recent boom in development of CRISPR-Cas-based technologies (55). For example, it has been proposed that anti-CRISPR genes can be used as regulatory tools to improve the use of CRISPR-Cas9 systems to control gene expression in various organisms by preventing off-target editing (56). Additionally, we envisage that anti-CRISPR genes can be powerful tools in the fight to control multi-drug resistant bacteria by providing virulent bacteriophages engineered for therapy purposes with a repertoire of genes enabling them to evade the CRISPR-Cas defence of the target pathogen and increase their host-range. To this end, it is necessary to not only continue discovering new anti-CRISPRs, but to thoroughly characterise them and identify their optimal versions for biotechnological use. We consider that a strategy such as the one presented in this study, i.e. combining genomics and phylogenomics analyses, *in vitro* evolution and functional characterisation, is a first step in that direction.

DATA AVAILABILITY

All sequences analysed in this work are publicly available from anti-CRISPRdb or GenBank. The corresponding accessions are indicated throughout the text and in figures 2, 3 and 4.

ACKNOWLEDGEMENTS

We thank Guadalupe Aguilar González from Unidad de Ácidos Nucleicos, Department of Genetics and Molecular Biology, and Dr Dulce Delgadillo Álvarez from Unidad de Genómica, Proteómica y Metabolómica - CINVESTAV, for technical assistance in Sanger sequencing and Miguel Ángel Moreno Galeana and Dr Eva Jacinto from DGMB, CINVESTAV, for technical support with lab experiments. We also thank Dr Herminia Loza Tavera from the Faculty of Chemistry, UNAM, for providing us with pUCP24 plasmid, and Prof Alan Davidson for sharing with us the strains PA14 WT, PA14 ΔCR and the phage JBD18.

FUNDING

W.F. acknowledges funding from National Council of Science and Technology-CONACYT (591274 and 706017) and Cambridge Trust (10469474). A.C. has been supported by the EMBL-EBI/Wellcome Trust Sanger Institute Join Post-Doctoral Fellowship Program (ESPOD). D.C was supported by the PhD scholarship 586079 from National Council of Science and Technology-CONACYT. G.G. acknowledges funding from National Council of Science and Technology-CONACYT CB 255255. Funding for open access charge: University of Cambridge.

AUTHORS CONTRIBUTION

W.F. and A.C. conceptualised the study and drafted the manuscript with input from other authors. W.F. carried out the positive selection analysis, random and site-directed mutagenesis, functional experiments, Sanger sequencing, and analysis of the protein structure. A.C. performed the homologs search, phylogenetics analysis, sequences comparison and comparative analysis of flanking regions. D.C. contributed with the conceptualisation of the study, performed infection assays and helped drafting the manuscript. A.D.C. cloned and tested the function of G2 WT. G.G. and L.K. contributed by supervising the work and revising the manuscript. G.G. was responsible for funding acquisition.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

1. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, **359**.

2. Mojica,F.J.M., Díez-Villaseñor,C., García-Martínez,J. and Soria,E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.*, **60**, 174–182.

3. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007)

CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.

4. Brouns,S.J.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J.H., Snijders,A.P.L., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, **321**, 960–964.

5. Marraffini,L.A. and Sontheimer,E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.*, **11**, 181–190.

6. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.

7. Cady,K.C., White,A.S., Hammond,J.H., Abendroth,M.D., Karthikeyan,R.S.G., Lalitha,P., Zegans,M.E. and O'Toole,G.A. (2011) Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology*, **157**, 430–437.

8. van Belkum,A., Soriaga,L.B., LaFave,M.C., Akella,S., Veyrieras,J.-B., Barbu,E.M., Shortridge,D., Blanc,B., Hannum,G., Zambardi,G., *et al.* (2015) Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *MBio*, **6**, e01796–15.

9. Stern,A. and Sorek,R. (2011) The phage-host arms race: shaping the evolution of microbes. *Bioessays*, **33**, 43–51.

10. Bondy-Denomy,J., Pawluk,A., Maxwell,K.L. and Davidson,A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, **493**, 429–432.

11. Pawluk,A., Bondy-Denomy,J., Cheung,V.H.W., Maxwell,K.L. and Davidson,A.R. (2014) A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *MBio*, **5**, e00896.

12. Dong,C., Hao,G.-F., Hua,H.-L., Liu,S., Labena,A.A., Chai,G., Huang,J., Rao,N. and Guo,F.-B. (2018) Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Res.*, **46**, D393–D398.

13. Pinilla-Redondo,R., Shehreen,S., Marino,N.D., Fagerlund,R.D., Brown,C.M., Sørensen,S.J., Fineran,P.C. and Bondy-Denomy,J. (2020) Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. *Nat. Commun.*, **11**, 5652.

14. Trasanidou,D., Gerós,A.S., Mohanraju,P., Nieuwenweg,A.C., Nobrega,F.L. and Staals,R.H.J. (2019) Keeping crispr in check: diverse mechanisms of phage-encoded anti-crisprs. *FEMS Microbiol. Lett.*, **366**.

15. Bondy-Denomy,J., Garcia,B., Strum,S., Du,M., Rollins,M.F., Hidalgo-Reyes,Y., Wiedenheft,B., Maxwell,K.L. and Davidson,A.R. (2015) Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature*, **526**, 136–139.

16. Pawluk,A., Shah,M., Mejdani,M., Calmettes,C., Moraes,T.F., Davidson,A.R. and Maxwell,K.L. (2017) Disabling a Type I-E CRISPR-Cas Nuclease with a Bacteriophage-Encoded Anti-CRISPR Protein. *mBio*, **8**.

17. Hardouin,P. and Goulet,A. (2020) Diversity of molecular mechanisms used by anti-CRISPR proteins: the tip of an iceberg? *Biochem. Soc. Trans.*, **48**, 507–516.

18. Kim,I., Koo,J., An,S.Y., Hong,S., Ka,D., Kim,E.-H., Bae,E. and Suh,J.-Y. (2020) Structural and mechanistic insights into the CRISPR inhibition of AcrIF7. *Nucleic Acids Res.*, **48**, 9959–9968.

19. Maxwell,K.L., Garcia,B., Bondy-Denomy,J., Bona,D., Hidalgo-Reyes,Y. and Davidson,A.R. (2016) The solution structure of an anti-CRISPR protein. *Nat. Commun.*, **7**, 13134.

20. Cady,K.C., Bondy-Denomy,J., Heussler,G.E., Davidson,A.R. and O'Toole,G.A. (2012) The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol.*, **194**, 5728–5738.

21. Cazares,A., Mendoza-Hernández,G. and Guarneros,G. (2014) Core and accessory genome architecture in a group of *Pseudomonas aeruginosa* Mu-like phages. *BMC Genomics*, **15**, 1146.

22. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of*

Molecular Biology, **215**, 403–410.

23. Simossis,V.A. and Heringa,J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.*, **33**, W289–94.

24. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.

25. Letunic,I. and Bork,P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.

26. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

27. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

28. Jesus,T.F., Ribeiro-Gonçalves,B., Silva,D.N., Bortolaia,V., Ramirez,M. and Carriço,J.A. (2019) Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.*, **47**, D188–D194.

29. Guy,L., Kultima,J.R. and Andersson,S.G.E. (2010) genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, **26**, 2334–2335.

30. Jolley,K.A. and Maiden,M.C.J. (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**.

31. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.

32. Nguyen,L.-T., Schmidt,H.A., von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

33. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl.*, **2**, 28–33.

34. Green,R. and Rogers,E.J. (2013) Transformation of chemically competent *E. coli*. *Methods Enzymol.*, **529**, 329–336.

35. Choi,K.-H., Kumar,A. and Schweizer,H.P. (2006) A 10-min method for preparation of highly electrocompetent *Pseudomonas aeruginosa* cells: application for DNA fragment transfer between chromosomes and plasmid transformation. *J. Microbiol. Methods*, **64**, 391–397.

36. Novogrodsky,A., Tal,M., Traub,A. and Hurwitz,J. (1966) The enzymatic phosphorylation of ribonucleic acid and deoxyribonucleic acid. II. Further properties of the 5'-hydroxyl polynucleotide kinase. *J. Biol. Chem.*, **241**, 2933–2943.

37. Yang,J. and Zhang,Y. (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.*, **43**, W174–81.

38. Pawluk,A., Staals,R.H.J., Taylor,C., Watson,B.N.J., Saha,S., Fineran,P.C., Maxwell,K.L. and Davidson,A.R. (2016) Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat Microbiol*, **1**, 16085.

39. Marino,N.D., Zhang,J.Y., Borges,A.L., Sousa,A.A., Leon,L.M., Rauch,B.J., Walton,R.T., Berry,J.D., Joung,J.K., Kleinstiver,B.P., *et al.* (2018) Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science*, **362**, 240–242.

40. Labrou,N.E. (2010) Random mutagenesis methods for in vitro directed enzyme evolution. *Curr. Protein Pept. Sci.*, **11**, 91–100.

41. Latino,L., Essoh,C., Blouin,Y., Vu Thien,H. and Pourcel,C. (2014) A novel *Pseudomonas aeruginosa* bacteriophage, Ab31, a chimera formed from temperate phage PAJU2 and *P. putida* lytic phage AF: characteristics and mechanism of bacterial resistance. *PLoS One*, **9**, e93777.

42. Mahendra,C., Christie,K.A., Osuna,B.A., Pinilla-Redondo,R., Kleinstiver,B.P. and Bondy-Denomy,J. (2020) Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. *Nat Microbiol*, **5**, 620–629.

43. Briones,C. (2014) Evolution, In Vitro. *Encyclopedia of Astrobiology*, 10.1007/978-3-642-27833-4_1726-3.

44. Keefe,A.D. and Szostak,J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.

45. Liebeton,K., Zonta,A., Schimossek,K., Nardini,M., Lang,D., Dijkstra,B.W., Reetz,M.T. and Jaeger,K.E. (2000) Directed evolution of an enantioselective lipase. *Chem. Biol.*, **7**, 709–718.

46. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.

47. Bloom,J.D., Silberg,J.J., Wilke,C.O., Drummond,D.A., Adami,C. and Arnold,F.H. (2005) Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 606–611.

48. Besenmatter,W., Kast,P. and Hilvert,D. (2007) Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins*, **66**, 500–506.

49. Hom,N., Gentles,L., Bloom,J.D. and Lee,K.K. (2019) Deep Mutational Scan of the Highly Conserved Influenza A Virus M1 Matrix Protein Reveals Substantial Intrinsic Mutational Tolerance. *J. Virol.*, **93**.

50. Cote-Hammarlof,P.A., Fragata,I., Flynn,J., Mavor,D., Zeldovich,K.B., Bank,C. and Bolon,D.N.A. (2021) The Adaptive Potential of the Middle Domain of Yeast Hsp90. *Mol. Biol. Evol.*, **38**, 368–379.

51. Gabel,C., Li,Z., Zhang,H. and Chang,L. (2021) Structural basis for inhibition of the type I-F CRISPR–Cas surveillance complex by AcrIF4, AcrIF7 and AcrIF14. *Nucleic Acids Research*, **49**, 584–594.

52. Callaway,E. (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, **588**, 203–204.

53. Pace,C.N., Horn,G., Hebert,E.J., Bechert,J., Shaw,K., Urbanikova,L., Scholtz,J.M. and Sevcik,J. (2001) Tyrosine hydrogen bonds make a large contribution to protein stability. *J. Mol. Biol.*, **312**, 393–404.

54. Zhu,B.Y., Zhou,N.E., Kay,C.M. and Hodges,R.S. (1993) Packing and hydrophobicity effects on protein folding and stability: effects of beta-branched amino acids, valine and isoleucine, on the formation and stability of two-stranded alpha-helical coiled coils/leucine zippers. *Protein Sci.*, **2**, 383–394.

55. Donohoue,P.D., Barrangou,R. and May,A.P. (2018) Advances in Industrial Biotechnology Using CRISPR-Cas Systems. *Trends Biotechnol.*, **36**, 134–146.

56. Marino,N.D., Pinilla-Redondo,R., Csörgő,B. and Bondy-Denomy,J. (2020) Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nat. Methods*, **17**, 471–479.