

Article

Comparative Analysis and Phylogenetic Insights of Cas14-Homology Proteins in Bacteria and Archaea

Numan Ullah, Naisu Yang, Zhongxia Guan, Kuilin Xiang, Yali Wang, Mohamed Diaby, Cai Chen, Bo Gao  and Chengyi Song * 

College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, China; dh18056@stu.yzu.edu.cn (N.U.); dx120180101@yzu.edu.cn (N.Y.); mx120190652@yzu.edu.cn (Z.G.); mx120220867@stu.yzu.edu.cn (K.X.); dx120180099@yzu.edu.cn (Y.W.); dh18035@yzu.edu.cn (M.D.); 007302@you.edu.cn (C.C.); bgao@yzu.edu.cn (B.G.)

* Correspondence: cysong@yzu.edu.cn

Abstract: Type-V-F Cas12f proteins, also known as Cas14, have drawn significant interest within the diverse CRISPR-Cas nucleases due to their compact size. This study involves analyzing and comparing Cas14-homology proteins in prokaryotic genomes through mining, sequence comparisons, a phylogenetic analysis, and an array/repeat analysis. In our analysis, we identified and mined a total of 93 Cas14-homology proteins that ranged in size from 344 aa to 843 aa. The majority of the Cas14-homology proteins discovered in this analysis were found within the *Firmicutes* group, which contained 37 species, representing 42% of all the Cas14-homology proteins identified. In archaea, the *DPANN* group had the highest number of species containing Cas14-homology proteins, a total of three species. The phylogenetic analysis results demonstrate the division of Cas14-homology proteins into three clades: Cas14-A, Cas14-B, and Cas14-U. Extensive similarity was observed at the C-terminal end (CTD) through a domain comparison of the three clades, suggesting a potentially shared mechanism of action due to the presence of cutting domains in that region. Additionally, a sequence similarity analysis of all the identified Cas14 sequences indicated a low level of similarity (18%) between the protein variants. The analysis of repeats/arrays in the extended nucleotide sequences of the identified Cas14-homology proteins highlighted that 44 out of the total mined proteins possessed CRISPR-associated repeats, with 20 of them being specific to Cas14. Our study contributes to the increased understanding of Cas14 proteins across prokaryotic genomes. These homologous proteins have the potential for future applications in the mining and engineering of Cas14 proteins.

Keywords: Cas14; Cas proteins; Cas14-homology protein; Cas14 mining; phylogenetics



Citation: Ullah, N.; Yang, N.; Guan, Z.; Xiang, K.; Wang, Y.; Diaby, M.; Chen, C.; Gao, B.; Song, C. Comparative Analysis and Phylogenetic Insights of Cas14-Homology Proteins in Bacteria and Archaea. *Genes* **2023**, *14*, 1911. <https://doi.org/10.3390/genes14101911>

Received: 9 September 2023

Revised: 29 September 2023

Accepted: 3 October 2023

Published: 6 October 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated Proteins (Cas) are a part of the bacterial and archaeal defense system. These proteins have the capability to eliminate invading viruses and mobile genetic elements (MGEs) and have also been used in gene editing [1,2]. In the past decade, researchers have primarily relied on Cas proteins that have over 1000 amino acids, such as Cas9 and Cas12a, for gene-editing purposes [3–5]. However, these longer proteins have encountered various challenges, such as limitations in delivery and low efficiency [6]. To address these challenges, miniature Cas proteins, such as CasΦ, Cas13, and Cas14, have been developed for gene editing and regulatory purposes. These proteins have a size between approximately 400 and 700 amino acids and consist of the RuvC nuclease domain [7–11]. The development of these smaller proteins holds significant importance for gene therapy applications as they can be encapsulated within FDA-approved adenoviral vectors (AVV) thanks to their smaller size [12].

Cas14 (Cas12f), a member of the class-2 type-V-F CRISPR-associated effector nuclease family, has been discovered through an analysis of genomic and metagenomic databases [13]. By conducting a comprehensive analysis of extensive terabase-scale sequencing datasets, the group uncovered a putative group of single-effector Cas proteins. This process involved meticulously sifting through the datasets to identify genes with a RuvC-like domain located proximal to CRISPR loci. The resulting collection of identified proteins signifies a significant advancement in the field due to their highly novel nature. Phylogenetically, Cas14 shares similarities with the bacterial RuvC-containing proteins C2c10 and C2c9, which are typically found near Cas arrays rather than Cas genes [13]. Initially discovered in uncultured archaea, these miniature effectors are capable of making RNA-guided single-strand (ssDNA) breaks in the target DNA without the need for any PAM sequence [13]. However, later studies have demonstrated that these effectors can induce double-stranded breaks (DSB) using the T-rich PAM sequence [8,14,15]. The majority of identified Cas14 proteins were ineffective in shielding *Escherichia coli* from invading double-stranded DNA. Out of the ten Cas14 proteins tested, only AsCas12f and SpCas12f exhibited the ability to mediate plasmid interference in *E. coli* [14–16]. Several studies have recently focused on gRNA engineering approaches in order to utilize Cas14 as a tool for editing genes in mammalian cells. Although several Cas14 proteins were found to cleave double-stranded target DNA in bacteria, no detectable activity in mammalian cells has been reported. Through the optimization of guide gRNAs and multiple rounds of protein engineering and screening, Cas12f variants were created and given the name CasMINI [17]. These variants demonstrated improved activity in the mammalian genome. Furthermore, not only have the biochemical features of SpCas12f been deciphered, but its effectiveness as a gene-editing tool in plants and human cells has also been demonstrated [9]. Similarly, subsequent gRNA engineering in SpCas12f has improved its efficiency in human cells. With the engineered gRNA, SpCas12f was able to achieve editing efficiencies comparable to that of Cas12a, thus expanding the CasMINI toolbox [17,18].

The Cas14 protein structure was determined through X-ray crystallography [18,19]. The study revealed that the Cas12f protein is a homodimer with two identical subunits comprising a nuclease domain and a helicase-nuclease domain. The nuclease domain of Cas12f encompasses the active site for DNA cleavage, consisting of two metal ions: a magnesium ion and a manganese ion. The helicase-nuclease domain is responsible for unwinding the target DNA and stabilizing the protein-DNA complex.

Bioinformatics tools have played a crucial role in the discovery and development of CRISPR-Cas as a gene-editing technology [20]. Recent studies have focused on discovering and characterizing novel miniature type-V Cas12f nucleases that exhibit diverse protospacer adjacent motif (PAM) preferences [21]. A comparative phylogenetic analysis of Cas proteins is important for understanding the diversity and evolution of CRISPR-Cas systems, as well as discovering novel Cas proteins that have potential applications in gene editing and other biotechnological fields [22,23]. Bioinformatics tools, such as BLAST, MUSCLE alignment algorithm, and neighbor-joining consensus trees based on the Jukes–Cantor model, can be utilized for performing comparative phylogenetic analyses of Cas proteins [24,25]. These analyses aid in the classification of CRISPR-Cas systems into distinct classes, types, and subtypes based on the conservation of Cas protein sequences and the architectural features of Cas loci. Additionally, a comparative phylogenetic analysis aids in the identification of novel Cas proteins by evaluating the preservation of Cas protein sequences. Furthermore, it can be employed to reconstruct the ancestral gene content and track the gain and loss of genes during the evolution of Cas proteins [26,27]. With the extensive diversity of CRISPR-Cas genes in the prokaryotic genome and the growing number of genomic and metagenomic sequencing data of bacteria and archaea being submitted to public databases, manually identifying, classifying, and tracking their evolutionary background is impractical [27,28]. Therefore, various automated approaches have been developed to enhance the gene-editing toolbox of Cas proteins [29–34]. These bioinformatics programs can predict the presence

and location of CRISPR-Cas genes within a genome, as well as identify their specific class and subtype, shedding light on their evolutionary origins and potential functional roles.

Previous phylogenetic analyses of the Cas14 protein family have yielded the identification of multiple Cas14 variants. For instance, recent studies have led to the discovery of three novel Cas12f effectors, referred to as μ Cas, which were derived from metagenome-assembled genomes (MAGs) found in ruminant microbiomes. In their study, Kong et al. characterized six CRISPR-Cas12f1 systems and specifically chose OsCas12f1 and RhCas12f1 for further investigation. [35]. OsCas12f1 recognizes a 5' T-rich protospacer adjacent motif (PAM), while RhCas12f1 recognizes a 5' C-rich PAM. To enhance their editing efficiency and broaden the recognition range of PAMs, protein and sgRNA engineering techniques were employed to develop advanced variants of OsCas12f1 (enOsCas12f1) and RhCas12f1 (enRhCas12f1). These enhanced variants showed increased editing efficiency and wider PAM recognition compared to the engineered variant Un1Cas12f1 (Un1Cas12f1_ge4.1) [35]. Additionally, an inducible-enOsCas12f1 construct was developed by fusing the destabilized domain with enOsCas12f1. Through the delivery of a single adeno-associated virus, the *in vivo* activity of inducible-enOsCas12f1 was demonstrated. These findings highlight the ongoing exploration of the diverse repertoire of Cas proteins and their potential functions within microbial communities. [21] However, these studies have not identified any discernible distinctions among the various subtypes of Cas14. In this study, our goal was to elucidate the diversity of Cas14-homology proteins and to characterize the domain disparities within the Cas14 protein family. By thoroughly analyzing publicly available databases, we successfully retrieved a set of Cas14-homology proteins. Our findings demonstrated the domain organization, sequence similarity, species distribution, and repeat analysis of these proteins, offering novel insights into the intricate and heterogeneous structure of the Cas14 family. Moreover, mining Cas14-homology proteins holds the potential to serve as a valuable reference for further functional enhancement in Cas14 engineering.

2. Materials and Methods

2.1. The Cas14 Mining

All available viral and prokaryotic genomes were downloaded from the NCBI database. To generate protein sequences, the transeq program of Emboss software was utilized. This program is capable of translating in any of the three forward or three reverse sense frames or in all six frames. The Cas14 sequences were obtained from reliable references. The CTD of Cas14, which includes motifs of RuvC segment I, II, and III, as well as recognition lobe 2 (REC2), Lid, and Nuc (the target nucleic acid binding), were aligned to create a hidden Markov model (HMM) profile named as Cas14-CTD.hmm. This profile is provided in the supplementary file. The Cas14 sequences were initially searched using the hmmsearch program of HMMER3 [36] in the translated proteins with Cas14-CTD.hmm with an *e*-value of 10^{-10} . Then, the protein sequences with 300 amino acids upstream and 100 amino acids downstream flanks of the hmmsearch hits were extracted. These sequences were then filtered using the local BLASTP program against all available Cas14 sequences, selecting candidates with an *e*-value higher than 10^{-30} . Subsequently, the genomic sequences with 20 kb flanks of these candidates were extracted and submitted for the annotation of the Cas14 protein and CRISPR array using the CRISPRCasTyper program [37]. A size filter was applied to the obtained Cas14 protein, only retaining proteins with a length greater than 300 amino acids. The presence of the CTD was confirmed in the sequences using alignment with MAFFT. Finally, the putative Cas14 proteins (>300 amino acids) were subjected to further domain and phylogenetic analyses.

2.2. Domain and Phylogenetic Analysis

The obtained Cas14 subfamilies, together with the reference sequences of Cas14, were submitted for alignment using the E-INS-I method from the MAFFT software [38]. The phylogenetic tree was inferred by using the maximum likelihood method in the IQ-TREE

program [39], with a best-suited aa substitution model selected by ModelFinder, and the ultrafast bootstrap approach with 1000 replicates was applied.

2.3. CRISPR Array/Repeats Prediction

The sequences identified through the phylogenetic analysis as Cas14 were subjected to a CRISPR Array analysis. Briefly, the prediction and classification of Cas operons and their spacers were carried out by CRISPRCasTyper 1.2.4 (<https://github.com/Russel88/CRISPRCasTyper>, accessed on 5 August 2023) [37]. Briefly, the extended 20 kb (per side) nucleotide sequences were taken for the proteins identified as Cas14 through blastp, and subsequent phylogenetic analyses were carried out with the CRISPRCasTyper.

3. Results

3.1. Classification of Cas14-Homology Proteins

All the mined Cas14-homology proteins were submitted for the primary phylogenetic analysis, and the minor branches in the tree with less than four sequences were removed, which was implemented to exclude rare sequences. Then, all the sequences (93 sequences, Supplementary Table S1) that passed the threshold were used for the IQ-Tree reconstruction. The ultimate phylogenetic tree classified the mined Cas14 proteins, along with the Cas14 proteins from published data, into three primary branches (Cas14-A, Cas14-B, and Cas14U) with strong support, as evidenced by bootstrap values exceeding 70%. The branch Cas14U (26 sequences) seems to be the most prominent, while the remaining branches, Cas14A (58 sequences) and Cas14B (9 sequences), exhibit potential for further subdivision into smaller branches. Within Cas14A, there are two sub-branches named Cas14A-I and Cas14A-II. Analogously, Cas14B is segregated into two subbranches, Cas14B1 and Cas14B2 (see Figure 1). In the analysis, the Cas1 protein, which is a universally conserved component of the CRISPR prokaryotic immune defense system, served as an outgroup for tree rooting [40] (Figure 1).

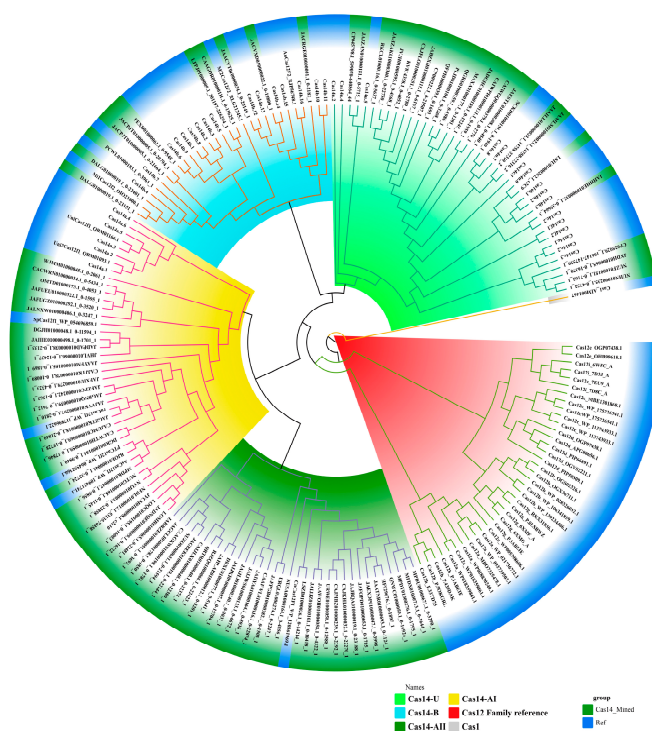


Figure 1. The phylogenetic tree of Cas14-homology proteins. The phylogenetic tree of mined Cas14 proteins with reference sequences from the Cas12a family and other RuvC-containing proteins. The Cas14-homology proteins were classified into three main branches (Cas14-AI, Cas14-AII, Cas14-B, and Cas14-U), as indicated with blue, purple, and orange ligands.

3.2. Domain Organization of Cas14-Homology Proteins

The majority of the Cas14-homology proteins we identified belong to the typical miniature Cas protein category, with lengths ranging from 344 aa to 843 aa. Varied sizes were observed among the Cas14-A, Cas14-B, and Cas14U clades, with Cas14-B having the longest average length (~590 aa) and Cas14-A being the shortest, with an average size of ~447 aa (Supplementary Table S1). The crystal structure of Cas14, which was recently determined [19], provides insight into its recognition and cleavage mechanism [41]. By aligning the mined Cas14 proteins with the reference Cas14 proteins, Un1Cas12f1, the domain organization of the three clades of Cas14 (Cas14-A, Cas14-B, and Cas14U) was inferred (Supplementary Figure S3). This analysis reveals that the NTD in all three clades contains a Rec domain sandwiched between two WED domains (Figure 2A). At the CTD of Cas14, three RuvC domains are found in conjunction with the Rec2, lid, and Nuc domains. Compared to the NTD, the CTD of all three clades exhibits a high degree of conservation. The conservation of the CTD in all three branches may be attributed to the presence of RuvC (I, II, III) segments, implying a potential shared excision mechanism among these proteins. Based on the alignment, some differences in the domain organization can be observed among the three clads. Notably, Cas14-B seems to have a different domain organization, especially at the NTD. Moreover, domain differences within the clades were also observed; for instance, Cas14-A1 and Cas14-II seem to have differences at the WED and REC domains at the NTD (Supplementary Figure S2).

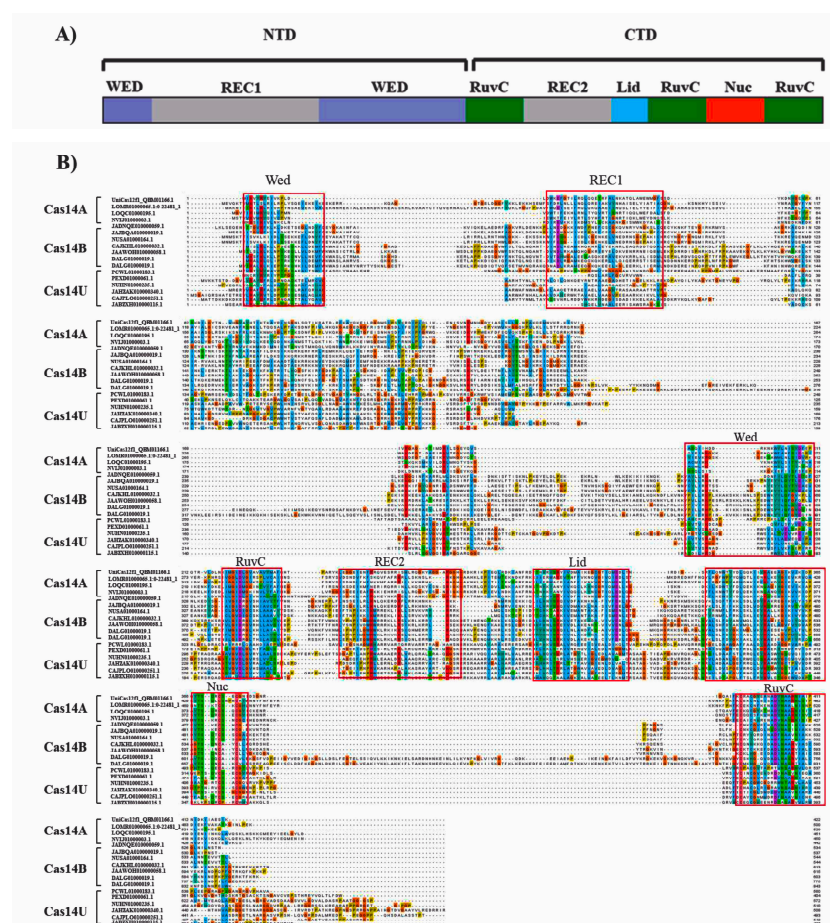


Figure 2. The alignments of Cas14-A, Cas14-B, and Cas14-U protein sequences. The alignment was performed with Mafft and drawn by Jalview Version 2. (A) Schematic of Cas14. The key domains include WED (purple), REC domain (gray), RuvC I, II, III (green), Lid (blue), and Nuc (orange). (B) The alignment of several sequences selected from Cas14-A, Cas14-B, and Cas14-U protein subgroups were identified based on the phylogenetic tree.

The overall sequence identity of the Cas14-homology proteins analyzed in this study was generally low, ranging from 6% to 70%, with an average identity of 18% (Figure 3A). Furthermore, a more detailed analysis of the sequence identity within each clade revealed that the Cas14-U clade had a sequence identity range of 7% to 70% and an average identity of 27% (Supplementary Figure S4). In terms of the specific clades, the Cas14A clade exhibited an average sequence identity of 26%, ranging from 13% to 99%, while the Cas14-B clade displayed an average sequence identity of 36%, ranging from 10% to 97% (Supplementary Figure S4).

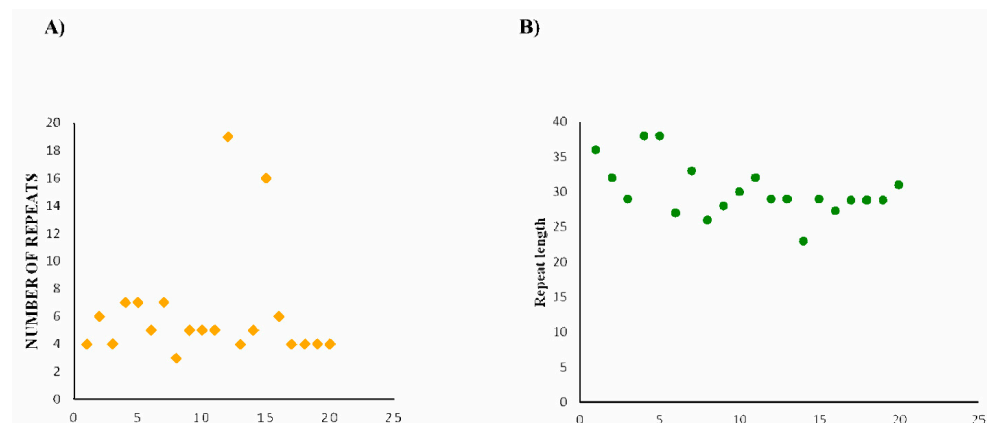


Figure 3. Distribution of the number of repeats and repeat length in the nucleotide sequences of mined Cas14 proteins with arrays. (A) The number of repeats in each nucleotide sequence. (B) The length (in bp) of repeats for each of the nucleotide sequences in which the array was detected (refer to Supplementary Table S2).

3.3. Putatively Functional Cas14 Proteins

The genomic coordinates flanking the investigated Cas14-homology proteins were screened using the standalone version of CRISPR-CasTyper to identify the presence of CRISPR arrays/repeats. A 20 kb nucleotide sequence on each side of the target region was subjected to analysis. The results revealed the existence of 44 sequences with arrays (as presented in Supplementary Table S2 and Supplementary Data), with 20 of these sequences being classified as type-V-F CRISPR-Cas (refer to Table 1). Notably, the data showcased a diverse range in the size of the repeat-spacer arrays, varying from 3 to 19 spacers (as summarized in Supplementary Table S2 and Supplementary Data). The CRISPR systems exhibited an average of nine repeats per locus. Additionally, the length of the repeat sequences was analyzed and displayed an average of 30 bp, ranging from 23 bp to 38 bp (as exemplified in Figure 3). Furthermore, an alignment of the putative Cas14 protein indicated that these putative proteins share major domains, particularly the RuvC domain that functions as an endonuclease domain that is present in the Cas14 reference protein UniCas12f1 (refer to Figure 4). Although this observation suggests that our identified homology proteins are putatively functional, it is important to note that Cas proteins require substrate recognition, inhibition, and excision, which involve multiple domains and neighboring arrays. Therefore, a thorough comparative functional comparison may necessitate further analysis. In contrast to the other Cas14-homology proteins identified in our study, the putative Cas14-homology proteins possess conserved domains both at the CTD and NTD. This suggests that these proteins may potentially be functionally active and further supports the presence of Cas14-associated arrays in our investigation.

Table 1. Predicted repeats from the mined Cas14 sequences.

Sequence ID	Consensus Repeat	Repeat Subtype
BJOE01000041.1:0-32724_1	CTCCAAACAGAATCATGCTTCTATGACTGTTCCGAG	V-F1
CAJUMC010000069.1:0-15728_7	CTTACACCATATACCTACGCATAGTTCGAGTC	V-F1
CP009222.1:0-21507_8	GTTCTTCCACACGCACACGAAGAAGATCCC	V-F2
DALG01000019.1:0-22401_10	AGTTGCATCTCTCATCTCGTTAATTCGTGCGCTGAAAC	V-F1
DALG01000019.1:0-23151_11	AGTTGCATCTCTCATCTCGTTAATTCGTGCGCTGAAAC	V-F1
DBXU01000077.1:0-5141_12	GCTGTGACTCATAGCAAAAAAGAAGGT	V-F1
DGRR01000184.1:0-9840_14	GATTATATCTGCTTGTATGGGTATACTGCGAGA	V-F1
HF929676.1:0-1897_15	TACACACTACATAGTCATTATATAAC	V-F1
JADGHC010000139.1:0-12245_16	GGGACTTCCCCGACGCGAGGACGACGG	V-F2
JADPAD010000038.1:0-2133_17	GTTTAAGAATAACAATAGTTGTATTTAAAT	V-F1
JAFXIU010000279.1:0-4323_19	GTTGCAACACGCGCATAAGGATGACTTGAAGG	V-F1
MPDK01000047.1:0-5856_23	GTTCACTCCACAAGCTAGCTCGCAAAC	V-F1
NUSA01000164.1:0-4856_24	GTTTTGAATAAACTATGTAGAATGTGAAT	V-F1
NVIJ01000003.1:0-22858_25	ATTTAAATACATCTTATGTTAGT	V-F1
LSZB01000026.1:0-14214_33	ATTTACATTTACATAGTTAAACTAAAAC	V-F1
JAHZAK010000340.1:0-22750-1_36	GTGTTCCCCGTATGTGCGGGGGTGAGC	V-F2
LOQC01000195.1:0-16003-1_48	ATTTCAATACATCTATTGTTATGTTTTAAC	V-F1
JADNQE010000059.1:0-16522-1_52	ATTTCAATACATCTATTGTTATGTTTTAAC	V-F1
LOMR01000065.1:0-22481-1_53	ATTTCAATACATCTATTGTTATGTTTTAAC	V-F1
NFDL01000012.1:53135-74358-1_58	AGGAAAAACATAATAATAGATGTATTGAAAT	V-F1
BJOE01000041.1:0-32724_1	CTCCAAACAGAATCATGCTTCTATGACTGTTCCGAG	V-F1
CAJUMC010000069.1:0-15728_7	CTTACACCATATACCTACGCATAGTTCGAGTC	V-F1
CP009222.1:0-21507_8	GTTCTTCCACACGCACACGAAGAAGATCCC	V-F2
DALG01000019.1:0-22401_10	AGTTGCATCTCTCATCTCGTTAATTCGTGCGCTGAAAC	V-F1

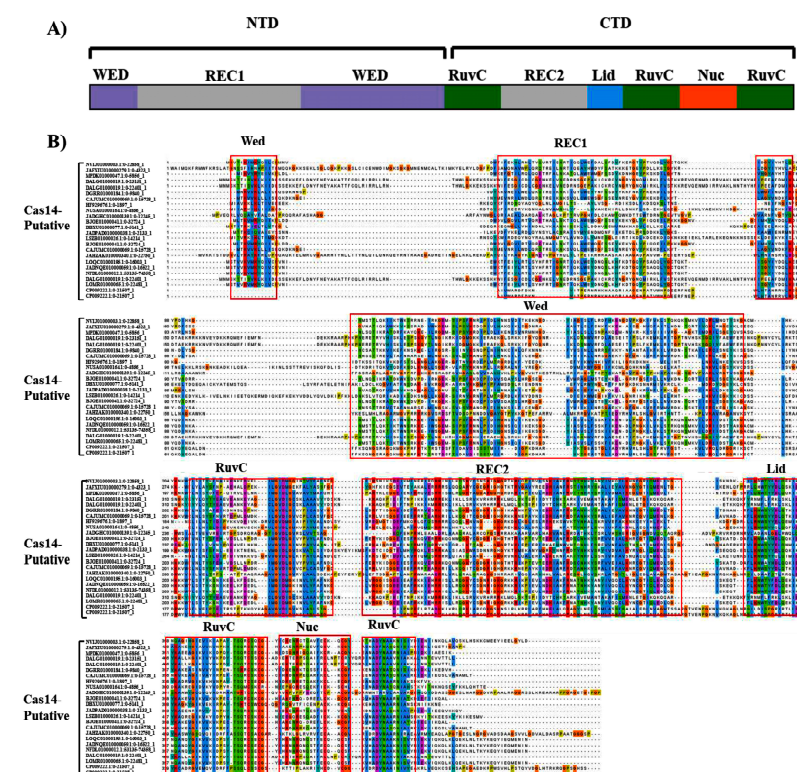


Figure 4. The alignments of putative active Cas14-homology protein sequences. The alignment was performed with MAFFT and drawn by Jalview Version 2. (A) Schematic of Cas14 (UniCas12f1). The key domains include the WED (grey), Rec1(purple), RuvC I, II, III (green), lid (blue), and Nuc (orange). (B) The alignment of Cas14 putative sequences with UniCas12f1. The WED, REC1, REC2, NUC, LID, and RuvC I, II, and III domains are marked using a red box.

3.4. Distribution of Cas14-Homology Proteins in Bacteria and Archaea

Our analysis revealed that mined Cas14 is present in both bacterial and archaeal species (Table 2). The *Firmicutes* group had the highest number of mined Cas14 homology proteins, with 37 species accounting for 42% of the total mined Cas14-homology proteins. The *Actinobacteria* group accounted for the second-highest number of species with mined Cas14, representing 18% of the total. Among the archaeal groups, there were a total of 10 species found to have Cas14 homology proteins. The *DPANN* group had the highest number of species with mined Cas14, representing 90% of the archaeal species having the protein. It accounted for 10% of all species in our analysis with mined Cas14-homology proteins (Table 2).

Table 2. Distribution of the number of Cas14 variants and putative Cas14 in bacterial and archaeal groups.

Group		All	Putative
Bacteria	Nitrospirae	0	0
	Fibrobacteres	0	0
	FCB group	14	1
	Bacteroidetes	0	0
	Chlorobi	0	0
	Gemmatimonadetes	0	0
	PVC group	0	0
	Verrucomicrobia	0	0
	Planctomycetes	0	0
	Chlamydiae	0	0
	Deinococcus-Thermus	0	0
	Firmicutes	37	16
	Armatimonadetes	0	0
	Terrabacteria group	0	0
	Chloroflexi	0	0
	Actinobacteria	16	3
	Candidatus	0	0
	Melainabacteria	0	0
	Cyanobacteria	1	1
	Candidatus	0	0
	Eremiobacteraeota	0	0
	Gammaproteobacteria	0	0
	Alphaproteobacteria	2	0
	Betaproteobacteria	1	0
	Proteobacteria	0	0
	unclassified	0	0
	environmental samples	0	0
	delta/epsilon subdivisions	0	0
	Zetaproteobacteria	0	0
	Oligoflexia	0	0
	Acidithiobacillia	0	0
	Candidatus	0	0
	Lambdaproteobacteria	0	0
	Candidatus	0	0
	Muproteobacteria	0	0

Table 2. Cont.

Group		All	Putative
Bacteria	Hydrogenophilalia	0	0
	Aquificae	0	0
	Thermotogae	0	0
	Deferribacteres	0	0
	Chrysiogenetes	0	0
	Thermodesulfobacteria	0	0
	Spirochaetes	0	0
	Fusobacteria	0	0
	Acidobacteria	0	0
	Dictyoglomi	0	0
	Calditrichaeota	0	0
	Nitrospinae/Tectomicrobia group	0	0
	Krumholzibacteriota	0	0
	Caldiserica/Crysericota group	0	0
	Coprothermobacteria	0	0
	Elusimicrobia	0	0
	Synergistetes	0	0
	unclassified Bacteria	0	0
	environmental samples	0	0
Archaea	Asgard group	0	0
	Candidatus Thermoplasmatota	0	0
	DPANN group	9	3
	Euyarchaeota	1	0
	TACK group	0	0

3.5. Distribution of Putatively Functional Cas14 Proteins

Our analysis uncovered a prominent distribution of putative Cas14 among various species. The results demonstrate that the *Firimutes* clade comprises the majority of species with putative Cas14 genes, representing 16% of all species in this category. Notably, within the archaeal clades, the *DPANN* group exhibits the highest number of species that harbor putative Cas14, accounting for 12% of all species with this gene (Table 2).

4. Discussion

Among the diverse CRISPR-Cas systems, class 2 is the most versatile due to its utilization of a single-effector protein. Particularly, miniature Cas proteins have gained significant importance in the field of gene therapy due to their smaller size in contrast to the commonly utilized Cas proteins such as Cas9 and Cas12a. The recent achievement in mining and gRNA engineering of miniature Cas12, specifically Cas14 (Cas12f), which is smaller than any other Cas proteins discovered thus far, has sparked a growing interest in discovering and characterizing these compact Cas proteins. Consequently, numerous studies have been conducted to enhance our comprehension of their structure, cleavage, and binding and compare the structure of Cas14 with its closely related counterpart in the

Cas12 family [19,41]. For example, a detailed comparison has been conducted between Cas14 and Cas Φ to compare their characteristics [42]. The comparison indicates that Cas14 and Cas12j, due to their smaller size, are less efficient than larger Cas endonucleases in cleaving target DNA. Their smaller size leads to slower cleavage kinetics and reduces genome-editing outcomes in human cells, as evidenced by multiple studies [14,17,43] and confirmed experimentally in a recent study [44]. This may be due to the absence of stabilizing contacts between the smaller proteins and the RNA–DNA heteroduplex. However, it has been demonstrated that modifying the Cas12f structure to enhance these contacts increases its efficiency in gene activation and editing [17,33]. Furthermore, the slower kinetics of the miniature Cas12 proteins may also impact their specificity, given that these proteins are designed to exhibit a certain degree of non-specificity towards their target DNA. However, preliminary reports indicate that they may demonstrate a limited tolerance for mismatches in the PAM-proximal region of the target [35].

The crystal structure of Cas14, which elucidates its recognition and cleavage mechanism, was recently determined [18,30]. According to the first study, the Cas12f protein is a homodimer with two identical subunits, each containing a nuclease domain and a helicase–nuclease domain [18]. The NTD consists of the wedge (WED), REC, and zinc finger (ZF) domains. In contrast, the CTD comprises the RuvC domain and a ZF domain referred to as the target nucleic-acid-binding (TNB) domain. The Cas12f dimer adopts a lobed structure, with a REC lobe and a nuclease (NUC) lobe, and the guide RNA–target DNA complex resides within the channel between them. The REC lobe is formed by the WED.1/ZF.1/REC.1 and WED.2/ZF.2/REC.2 domains of Cas12f.1 and Cas12f.2, respectively, whereas the NUC lobe consists of the RuvC.1/TNB.1 and RuvC.2/TNB.2 domains of Cas12f.1 and Cas12f.2.

Despite its smaller size, Cas12f encompasses all the typical domains found in other Cas12 proteins [41]. For instance, it possesses several functional domains, including RuvC, WED, and REC1. Among these domains, the RuvC domain plays a crucial role of excision. In Cas12f, substrate recognition activates only one RuvC domain in the dimer, capturing the substrate within the structure. This unique feature distinguishes Cas12f from other CRISPR nucleases. [45]. Upon substrate recognition, the RuvC domain, responsible for cleaving the target DNA, undergoes a close-to-open transition in the lid motif. The WED domain is composed of a seven-stranded β -barrel flanked by an α helix and a β hairpin. It adopts an oligonucleotide/oligosaccharide-binding fold similar to that of other nucleic acid-binding proteins [46]. Though the exact function of the WED domain in Cas12f remains unclear, it is believed to participate in the recognition and binding of target DNA. Additionally, the WED domain may play a role in the conformational changes that arise during substrate recognition and cleavage by the RuvC domain. A comparison of Cas14 with the compact Cas12 protein suggested that the closest relative to Cas12f is Cas12g, which consists of 767 amino acids and belongs to branch 3 of type-V nucleases, as determined by a phylogenetic analysis [41].

The Cas12f monomer is at the NTD, containing the REC1 and WED domains, and the C-terminal holds the RuvC, REC2 (included as part of the RuvC domain in [19]), and Nuc domains (TNB domain in [19]). The primary difference between the two is the REC1 domain, which can be broken down into two subdomains: REC1N (referred to as a zinc finger or ZF domain in [19]), which features a zinc finger with a zinc ion coordinated by four cysteines (C475, C478, C500, and C503), and REC1C, a three-helix bundle that acts as the main dimerization interface of Cas12f.

The majority of Cas14-mined proteins, except for AsCas12f1, UniCas12f1, and Sp-Cas12f, did not exhibit any DNA excision activity. These three proteins have been the focus of studies utilizing gRNA engineering to enhance their cutting efficiency in mammalian gene editing [34]. In a recent study, a comprehensive comparison was conducted to evaluate the efficiency and safety of Cas12f proteins, including AsCas12f1, CasMini, CasMINI_ge4.1, and Uni1Cas12f1_ge4.1, in comparison to commonly used Cas9, LbCas12a, and AsCas12a. The results showed that Cas12f nucleases demonstrated robust cleavage at the majority of

tested sites, with deletional fragments being the predominant outcome. In contrast, Cas9 and Cas12a exhibited comparatively higher editing efficacy across most tested sites. Importantly, cells edited with Cas12f nucleases showed minimal off-target hotspots, while cells edited with Cas9 and Cas12a exhibited observable hotspots. Additionally, Cas12f nucleases reduced the occurrence of chromosomal translocations, large deletions, and integrated vectors by 2–3 fold, as compared to Cas9 and Cas12a [44]. Moreover, recently there have been advancements in utilizing the smaller-sized Cas12f to develop miniature cytosine base editors (miniCBEs) and adenine base editors (miniABEs). These newly designed miniCBEs and miniABEs have demonstrated their efficiency in correcting pathogenic mutations in cell lines. Additionally, through the delivery of an adeno-associated virus, they have successfully introduced genetic mutations in the brain in vivo. These findings highlight the potential of engineered miniCBEs and miniABEs as effective tools in gene editing for various applications. However, further research is necessary to determine the optimal conditions for their use and to thoroughly evaluate their safety and efficacy in vivo [46].

This study explores Cas14-homology proteins in prokaryotes and classifies them into three groups: Cas14-A, Cas14-B, and Cas14-U, following phylogenetic analysis. The sequence comparisons reveal high similarity at the CTD among these groups, suggesting that they may have similar modes of action. In contrast, the overall sequence similarity among all mined Cas14 sequences is only 18%. Additionally, the analysis of repeats in the CRISPR arrays of the mined Cas14 proteins reveals a wide range of repeat numbers and lengths. These results enhance our understanding of the diversity and characteristics of Cas14 proteins. The successful classification of the mined Cas14-homology proteins into three distinct clades (Cas14-A, Cas14-B, and Cas14-U) offers valuable insights for future studies and the potential application of these proteins. The identification of high similarity at the CTD end among the three clades potentially leads to more efficient and targeted gene editing. However, the low overall sequence similarity among all mined Cas14 sequences underscores the necessity for further studies to comprehensively understand the molecular basis of the differences among these proteins. This information can yield valuable insights into the specificity and efficiency of Cas14-mediated gene editing and holds implications for the evolution and adaptation of the CRISPR-Cas system in various organisms, potentially being utilized to develop novel gene-editing strategies.

5. Conclusions

The current study identified 93 Cas14-homology proteins within prokaryotic genomes. The mined Cas14-homology proteins were classified into three clades, namely Cas14-A, Cas14-B, and Cas14-U, based on a phylogenetic analysis. Although the key domains among the Cas14 clades were found to be conserved, specifically at the CTD, their overall sequence identities are low. Moreover, repeats associated with the mined Cas14-homology proteins were also analysed, and based on the classification of the repeats, a total of 20 putatively active proteins were found.

This study offers significant insights into the diversity and characteristics of Cas14-homology proteins and holds important implications for understanding the classification and mining of Cas14-homology proteins in proteins. Further research is necessary to comprehensively understand the molecular basis of the differences among these proteins and develop more effective gene-editing tools.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/genes14101911/s1>: Figure S1: Phylogenetic tree of Cas14 mined protein with reference Cas14 proteins; Figure S2: Alignment of representative sequence from each clade of mined Cas14 homology proteins with Un1Cas12f, Figure S3: Alignment of all mined Cas14 homology proteins with Un1Cas12f, Figure S4: Sequence identity of all mined Cas14 homology proteins and individual clades, Table S1: Length of the mined Cas14 homology proteins, Table S2: CRISPR repeats of the identified Cas14 mined proteins.

Author Contributions: N.U. created the figures and edited the table. N.U. and M.D. analyzed the sequences and collected data. N.Y., Z.G. and K.X. collected the sequences. M.D., C.C., Y.W. and N.U. revised the manuscript. B.G. and C.S. conceived the project and helped write the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded with grants from the National Natural Science Foundation of China (32271508 and 31671313) and the High-end Talent Support Program of Yangzhou University to Chengyi Song.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank all the authors for their suggestions and critical comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mali, P.; Esvelt, K.M.; Church, G.M. Cas9 as a Versatile Tool for Engineering Biology. *Nat. Methods* **2013**, *10*, 957–963. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Barrangou, R.L.E. CRISPR Rewrites the Future of Medicine. *Cris. J.* **2022**, *5*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Cong, L.; Ran, F.A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P.D.; Wu, X.; Jiang, W.; Marraffini, L.A.; et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **2013**, *339*, 819–823. [\[CrossRef\]](#)
4. Kleinstiver, B.P.; Sousa, A.A.; Walton, R.T.; Tak, Y.E.; Hsu, J.Y.; Clement, K.; Welch, M.M.; Horng, J.E.; Malagon-Lopez, J.; Scarfò, I.; et al. Engineered CRISPR–Cas12a Variants with Increased Activities and Improved Targeting Ranges for Gene, Epigenetic and Base Editing. *Nat. Biotechnol.* **2019**, *37*, 276–282. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Tu, M.; Lin, L.; Cheng, Y.; He, X.; Sun, H.; Xie, H.; Fu, J.; Liu, C.; Li, J.; Chen, D.; et al. A New Lease of Life: Fncpf1 Possesses DNA Cleavage Activity for Genome Editing in Human Cells. *Nucleic Acids Res.* **2017**, *45*, 11295–11304. [\[CrossRef\]](#)
6. Liu, W.; Li, L.; Jiang, J.; Wu, M.; Lin, P. Applications and Challenges of CRISPR-Cas Gene-Editing to Disease Treatment in Clinics. *Precis. Clin. Med.* **2021**, *4*, 179–191. [\[CrossRef\]](#)
7. Liu, J.J.; Orlova, N.; Oakes, B.L.; Ma, E.; Spinner, H.B.; Baney, K.L.M.; Chuck, J.; Tan, D.; Knott, G.J.; Harrington, L.B.; et al. CasX Enzymes Comprise a Distinct Family of RNA-Guided Genome Editors. *Nature* **2019**, *566*, 218–223. [\[CrossRef\]](#)
8. Xu, X.; Chemparathy, A.; Zeng, L.; Kempton, H.R.; Shang, S.; Nakamura, M.; Qi, L.S. Engineered Miniature CRISPR-Cas System for Mammalian Genome Regulation and Editing. *Mol. Cell* **2021**, *81*, 4333–4345. [\[CrossRef\]](#)
9. Bigelyte, G.; Young, J.K.; Karvelis, T.; Budre, K.; Zedaveinyte, R.; Djukanovic, V.; Van Ginkel, E.; Paulraj, S.; Gasior, S.; Jones, S.; et al. Miniature Type V-F CRISPR-Cas Nucleases Enable Targeted DNA Modification in Cells. *Nat. Commun.* **2021**, *10*, 957–963. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Freije, C.A.; Myhrvold, C.; Boehm, C.K.; Lin, A.E.; Welch, N.L.; Carter, A.; Metsky, H.C.; Luo, C.Y.; Abudayyeh, O.O.; Gootenberg, J.S.; et al. Programmable Inhibition and Detection of RNA Viruses Using Cas13. *Mol. Cell* **2019**, *76*, 826–837. [\[CrossRef\]](#)
11. Pausch, P.; Soczek, K.M.; Herbst, D.A.; Tsuchida, C.A.; Al-Shayeb, B.; Banfield, J.F.; Nogales, E.; Doudna, J.A. DNA Interference States of the Hypercompact CRISPR–CasΦ Effector. *Nat. Struct. Mol. Biol.* **2021**, *28*, 652–661. [\[CrossRef\]](#)
12. Doudna, J.A. The Promise and Challenge of Therapeutic Genome Editing. *Nature* **2020**, *578*, 229–236. [\[CrossRef\]](#)
13. Harrington, L.B.; Harrington, L.B.; Burstein, D.; Chen, J.S.; Paez-espino, D.; Ma, E.; Witte, I.P.; Cofsky, J.C.; Kyrpides, N.C.; Banfield, J.F.; et al. Programmed DNA Destruction by Miniature CRISPR-Cas14 Enzymes. *Science* **2018**, *362*, 839–842. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Karvelis, T.; Bigelyte, G.; Young, J.K.; Hou, Z.; Zedaveinyte, R.; Budre, K.; Paulraj, S.; Djukanovic, V.; Gasior, S.; Silanskas, A.; et al. PAM Recognition by Miniature CRISPR-Cas12f Nucleases Triggers Programmable Double-Stranded DNA Target Cleavage. *Nucleic Acids Res.* **2020**, *48*, 5016–5023. [\[CrossRef\]](#)
15. Kim, D.Y.; Chung, Y.; Lee, Y.; Jeong, D.; Park, K.H.; Chin, H.J.; Lee, J.M.; Park, S.; Ko, S.; Ko, J.H.; et al. Hypercompact Adenine Base Editors Based on Transposase B Guided by Engineered RNA. *Nat. Chem. Biol.* **2022**, *18*, 1005–1013. [\[CrossRef\]](#)
16. Okano, K.; Sato, Y.; Hizume, T.; Honda, K. Genome Editing by Miniature CRISPR/Cas12f1 Enzyme in *Escherichia coli*. *J. Biosci. Bioeng.* **2021**, *132*, 120–124. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wang, Y.; Wang, Y.; Pan, D.; Yu, H.; Zhang, Y.; Chen, W.; Li, F.; Wu, Z.; Ji, Q. Guide RNA Engineering Enables Efficient CRISPR Editing with a Miniature *Syntrophomonas palmitatica* Cas12f1 Nuclease. *Cell Rep.* **2022**, *40*, 111418. [\[CrossRef\]](#)
18. Takeda, S.N.; Nakagawa, R.; Okazaki, S.; Hirano, H.; Kobayashi, K.; Kusakizako, T.; Nishizawa, T.; Yamashita, K.; Nishimasu, H.; Nureki, O. Structure of the Miniature Type V-F CRISPR-Cas Effector Enzyme. *Mol. Cell* **2021**, *81*, 558–570. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Naeem, M.; Alkhnbashi, O.S. Current Bioinformatics Tools to Optimize CRISPR/Cas9 Experiments to Reduce Off-Target Effects. *Int. J. Mol. Sci.* **2023**, *24*, 6261. [\[CrossRef\]](#)
20. Sharrar, A.; Arake de Tacca, L.; Collingwood, T.; Meacham, Z.; Rabuka, D.; Staples-Ager, J.; Schelle, M. Discovery and Characterization of Novel Type V Cas12f Nucleases with Diverse Protospacer Adjacent Motif Preferences. *Cris. J.* **2023**, *6*, 350–358. [\[CrossRef\]](#) [\[PubMed\]](#)

21. Makarova, K.S.; Koonin, E.V. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol. Biol.* **2015**, *1311*, 47–75. [[CrossRef](#)] [[PubMed](#)]
22. York, A. Metagenomics: Mining for CRISPR-Cas. *Nat. Rev. Microbiol.* **2017**, *15*, 133. [[CrossRef](#)]
23. Cornwell, W.; Nakagawa, S. Phylogenetic Comparative Methods. *Curr. Biol.* **2017**, *27*, R333–R336. [[CrossRef](#)]
24. Jacques, F.; Bolivar, P.; Pietras, K.; Hammarlund, E.U. Roadmap to the Study of Gene and Protein Phylogeny and evolution—A Practical Guide. *PLoS ONE* **2023**, *18*, e0279597. [[CrossRef](#)] [[PubMed](#)]
25. Koonin, E.V.; Makarova, K.S. Origins and Evolution of CRISPR-Cas Systems. *Philos. Trans. R. Soc. B Biol. Sci.* **2019**, *374*, 20180087. [[CrossRef](#)]
26. Makarova, K.S.; Wolf, Y.I.; Iranzo, J.; Shmakov, S.A.; Alkhnbashi, O.S.; Brouns, S.J.J.; Charpentier, E.; Cheng, D.; Haft, D.H.; Horvath, P.; et al. Evolutionary Classification of CRISPR–Cas Systems: A Burst of Class 2 and Derived Variants. *Nat. Rev. Microbiol.* **2020**, *18*, 67–83. [[CrossRef](#)]
27. Makarova, K.S.; Koonin, E.V. Evolution and Classification of CRISPR-Cas Systems and Cas Protein Families. In *CRISPR-Cas Systems: RNA-Mediated Adaptive Immunity in Bacteria and Archaea*; Springer: Berlin/Heidelberg, Germany, 2013; ISBN 9783642346576.
28. Lange, S.J.; Alkhnbashi, O.S.; Rose, D.; Will, S.; Backofen, R. CRISPRmap: An Automated Classification of Repeat Conservation in Prokaryotic Adaptive Immune Systems. *Nucleic Acids Res.* **2013**, *41*, 8034–8044. [[CrossRef](#)] [[PubMed](#)]
29. Biswas, A.; Staals, R.H.J.; Morales, S.E.; Fineran, P.C.; Brown, C.M. CRISPRDetect: A Flexible Algorithm to Define CRISPR Arrays. *BMC Genom.* **2016**, *17*, 356. [[CrossRef](#)]
30. Crawley, A.B.; Henriksen, J.R.; Barrangou, R. CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems. *Cris. J.* **2018**, *1*, 171–181. [[CrossRef](#)] [[PubMed](#)]
31. Abby, S.S.; Néron, B.; Ménager, H.; Touchon, M.; Rocha, E.P.C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS ONE* **2014**, *9*, e110726. [[CrossRef](#)]
32. Couvin, D.; Bernheim, A.; Toffano-Nioche, C.; Touchon, M.; Michalik, J.; Néron, B.; Rocha, E.P.C.; Vergnaud, G.; Gautheret, D.; Pourcel, C. CRISPRCasFinder, an Update of CRISPRFinder, Includes a Portable Version, Enhanced Performance and Integrates Search for Cas Proteins. *Nucleic Acids Res.* **2018**, *46*, W246–W251. [[CrossRef](#)]
33. Chai, G.; Yu, M.; Jiang, L.; Duan, Y.; Huang, J. HMMCAS: A Web Tool for the Identification and Domain Annotations of CAS Proteins. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2019**, *16*, 1313–1315. [[CrossRef](#)]
34. Mitrofanov, A.; Alkhnbashi, O.S.; Shmakov, S.A.; Makarova, K.S.; Koonin, E.V.; Backofen, R. CRISPRidentify: Identification of CRISPR Arrays Using Machine Learning Approach. *Nucleic Acids Res.* **2021**, *49*, e20. [[CrossRef](#)]
35. Kong, X.; Zhang, H.; Li, G.; Wang, Z.; Kong, X.; Wang, L.; Xue, M.; Zhang, W.; Wang, Y.; Lin, J.; et al. Engineered CRISPR–OsCas12f1 and RhCas12f1 with Robust Activities and Expanded Target Range for Genome Editing. *Nat. Commun.* **2023**, *14*, 2046. [[CrossRef](#)]
36. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Res.* **2013**, *41*, e121. [[CrossRef](#)]
37. Russel, J.; Pinilla-Redondo, R.; Mayo-Muñoz, D.; Shah, S.A.; Sørensen, S.J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *Cris. J.* **2020**, *3*, 462–469. [[CrossRef](#)] [[PubMed](#)]
38. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)] [[PubMed](#)]
39. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [[CrossRef](#)]
40. Wiedenheft, B.; Zhou, K.; Jinek, M.; Coyle, S.M.; Ma, W.; Doudna, J.A. Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. *Structure* **2009**, *17*, 904–912. [[CrossRef](#)]
41. Xiao, R.; Li, Z.; Wang, S.; Han, R.; Chang, L. Structural Basis for Substrate Recognition and Cleavage by the Dimerization-Dependent CRISPR-Cas12f Nuclease. *Nucleic Acids Res.* **2021**, *49*, 4120–4128. [[CrossRef](#)] [[PubMed](#)]
42. Nguyen, G.T.; Dhingra, Y.; Sashital, D.G. Miniature CRISPR-Cas12 Endonucleases—Programmed DNA Targeting in a Smaller Package. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102466. [[CrossRef](#)]
43. Pausch, P.; Al-Shayeb, B.; Bisom-Rapp, E.; Tsuchida, C.A.; Li, Z.; Cress, B.F.; Knott, G.J.; Jacobsen, S.E.; Banfield, J.F.; Doudna, J.A. Crispr-Casf from Huge Phages Is a Hypercompact Genome Editor. *Science* **2020**, *369*, 333–337. [[CrossRef](#)] [[PubMed](#)]
44. Xin, C.; Yin, J.; Yuan, S.; Ou, L.; Liu, M.; Zhang, W.; Hu, J. Comprehensive Assessment of Miniature CRISPR-Cas12f Nucleases for Gene Disruption. *Nat. Commun.* **2022**, *13*, 5623. [[CrossRef](#)]
45. Tong, B.; Dong, H.; Cui, Y.; Jiang, P.; Jin, Z.; Zhang, D. The Versatile Type V CRISPR Effectors and Their Application Prospects. *Front. Cell Dev. Biol.* **2021**, *8*, 622103. [[CrossRef](#)] [[PubMed](#)]
46. Zhang, S.; Song, L.; Yuan, B.; Zhang, C.; Cao, J.; Chen, J.; Qiu, J.; Tai, Y.; Chen, J.; Qiu, Z.; et al. TadA Reprogramming to Generate Potent Miniature Base Editors with High Precision. *Nat. Commun.* **2023**, *14*, 413. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.