Taxonomic Distribution, Phylogenetic Relationship, and Domain Conservation of CRISPR-Associated Cas Proteins

Bioinformatics and Biology Insights
Volume 18: 1–9
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322241274961
Sage

Weerakkody Ranasinghe[1], Dorcie Gillette[2], Alexis Ho[1], Hyuk Cho[3] and Madhusudan Choudhary[1]

[1]Department of Biological Sciences, Sam Houston State University, Huntsville, TX, USA.
[2]Department of Surgery, The University of Iowa Hospitals and Clinics, Iowa City, IA, USA.
[3]Department of Computer Science, Sam Houston State University, Huntsville, TX, USA.

ABSTRACT: CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is a naturally occurring genetic defense system in bacteria and archaea. It is comprised of a series of DNA sequence repeats with spacers derived from previous exposures to plasmid or phage. Further understanding and applications of CRISPR system have revolutionized our capacity for gene or genome editing of prokaryotes and eukaryotes. The CRISPR systems are classified into 3 distinct types: type I, type II, and type III, each of which possesses an associated signature protein, Cas3, Cas9, and Cas10, respectively. As the CRISPR loci originated from earlier independent exposures of foreign genetic elements, it is likely that their associated signature proteins may have evolved rapidly. Also, their functional domain structures might have experienced different selective pressures, and therefore, they have differentially diverged in their amino acid sequences. We employed genomic, phylogenetic, and structure-function constraint analyses to reveal the evolutionary distribution, phylogenetic relationship, and structure-function constraints of Cas3, Cas9, and Cas10 proteins. Results reveal that all 3 Cas-associated proteins are highly represented in the phyla *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*, including both pathogenic and non-pathogenic species. Genomic analysis of homologous proteins demonstrates that the proteins share 30% to 50% amino acid identity; therefore, they are low to moderately conserved and evolved rapidly. Phylogenetic analysis shows that 3 proteins originated monophyletically; however, the evolution rates were different among different branches of the clades. Furthermore, structure-function constraint analysis reveals that both Cas3 and Cas9 proteins experiences low to moderate levels of negative selection, and several protein domains of Cas3 and Cas9 proteins are highly conserved. To the contrary, most protein domains of Cas10 proteins experience neutral or positive selection, which supports rapid genetic divergence and less structure-function constraints.

KEYWORDS: CRISPR, Cas-associated proteins, phylogenetic analysis, and structure-function constraint

RECEIVED: January 3, 2024. **ACCEPTED:** July 30, 2024.

TYPE: Research Article

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Madhusudan Choudhary, Department of Biological Sciences, Sam Houston State University, Huntsville, TX 77341, USA. Email: mchoudhary@shsu.edu

Introduction

The ability of clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 to easily program the system provides an exciting gene-editing technology capable of editing the genomes of both prokaryotes and eukaryotes, including the human genome. The CRISPR/Cas9 system has been successfully applied to many different types of cells and organisms, such as bacteria, fungi, viruses, parasites, plants, animals, and human cell lines.[1] In addition, the system has been successfully employed to create transgenic animals.[2]

The CRISPR is a naturally occurring genetic defense system in bacteria and archaea.[3] The CRISPR locus has a unique DNA sequence structure consisting of direct repeats, ranging from 21 to 37 nucleotides, interspaced by non-repetitive sequences of similar size.[4] The CRISPR defense has 3 separate phases: adaptation, expression, and interference. During adaptation, a short DNA fragment is removed from an invasive DNA and is incorporated into the CRISPR array in a site-specific manner to create a new spacer. In the second step, transcription of the CRISPR array results in a precursor CRISPR RNA (pre-crRNA) that binds with Cas proteins to undergo additional processing into mature crRNAs. Finally, during interference, the combined activity of crRNAs and Cas proteins recognizes and seeks the newly invasive DNAs and destroys the target nucleic acids.[5] The CRISPR systems are extensively distributed across the genomes of 42% of bacteria and 85% of archaea.[6] The CRISPR array matches the phage sequences that commonly invade bacteria. Previous studies corroborated that the non-repetitive spacers served as templates to target invading bacteriophages following previous exposures.[7-9] The CRISPR systems were dependent on DNA complementary pairing because if the spacer in the CRISPR locus was no longer complementary to the phage genome, the systems could not seek and destroy the newly infecting bacteriophages.[7] Furthermore, the CRISPR systems were discovered to be transferable from bacteria with naturally occurring CRISPR systems to those lacking a CRISPR system by horizontal gene transfer.[4] The CRISPR systems also require *cas* genes, which encode CRISPR-associated proteins for the functionality of the system.[8]

The CRISPR-Cas systems are classified into 2 major classes, class 1 and class 2, based on the number and protein composition participating in nucleic acid interference. These classes are further divided into 6 types and 33 sub-types, with multi-Cas protein effector complexes in Class 1 systems (types I, III, and IV) and a single effector protein in class 2 systems (types II, V, and VI).[9,10] The signature proteins for types I, II,

Table 1. Comparison of Cas proteins: Cas3, Cas9, and Cas10.

SIGNATURE PROTEIN	CLASS	TYPE	STRUCTURAL DOMAINS	FUNCTION AND MECHANISM	REFERENCES
Cas3	Class I	Type I	HD nuclease, 2 RecAs (RecA1 and RecA2), Linker, and C-terminal domain (CTD)	It encodes a cascade-like complex to recognize and destroy targets and it additionally requires PAM sequence.	11-13
Cas9	Class II	Type II	Three RuvCs (RuvCI, RuvCII, and RuvCIII), 3 RECs (2 REC1s and REC2), HNH, and PAM Interacting	It recognizes targets and destroys invasive DNAs during expression, and it requires trans-encoded small RNA (tracrRNA) and PAM sequence.	11,14-17
Cas10	Class I	Type III	HD nuclease, Cyclase/RRM, Zn Finger, D2, Palm/RRM, and Thumb	It recognizes target DNAs and RNAs, but mechanisms are poorly understood.	18-21

and III are Cas3, Cas9, and Cas10, respectively, and they perform different functions within their respective CRISPR systems as summarized in Table 1.

Cas3 is an ATP-dependent single-strand DNA (ssDNA) translocase/helicase that is linked to an HD-nuclease domain (histidine-aspartate [HD] nuclease domain) in many CRISPR systems. During the part of a process known as "CRISPR interference," the Cas3 translocase and nuclease activities break down DNA by reducing it to shorter fragments of tens of nucleotides, abolishing invading DNA.[12,22,23]

Cas9 is present in Eubacteria that cleaves viral DNA by unwinding and complementary pairing to the guide RNA and defend them against bacteriophages and plasmids.[5,24] Apart from that, Cas9 can recruit proteins to a target site enabling a powerful engineered sequence-specific gene-editing and gene regulation control mechanisms.[25,26] The multifunctional role of Cas9 is comprised of a more complex set of domains; however, the most important distinction is in its 2 nuclease domains, HNH and RuvC, which selectively cleave the target DNA. The dual role of Cas9 as a nuclease and an interferase lends to its simplicity for diverse genetic applications. The best characterized CRISPR system, type II CRISPR/Cas9, is a member of the class II system because it requires only the protein Cas9 for endonuclease activity.[27]

Cas10 proteins belonging to type III contain an N-terminal HD-nuclease domain, 2 PALM domains separated by a zinc-finger motif (ZF), and a C-terminal domain (CT).[28,29] It is thought that single-stranded DNase activity, observed for the HD domains of certain Cas10 proteins, promotes the nicking of ssDNA created during transcription.[30,31,32] Many Cas10 proteins lack an N-terminal HD domain, and the quantity and phylogenetic distribution of such truncated proteins have not been thoroughly evaluated.[6,33]

The high sequence similarity within type III CRISPR systems as well as their similarity with type I systems suggest that type I may have evolved from type III.[34] In addition, type I and III systems use homologous HD-nuclease domains for catalysis.[1,12,24] Furthermore, the multiprotein effector complexes of type I and III systems are more different from the single-protein activity of type II. The primary differences are due to the replacement of the HD-nuclease activity of the types I and III with the HNH and RuvC nuclease domains of the Cas9 protein of type II system.[22] The Cas9 protein of type II show no structural and functional similarity to any other proteins found in type I and III systems.[34] However, Cas9 appears to belong to a family of proteins that contains some protein family members in the type I system. It has been further indicated that the type II system originated from the fusion of the Cas9 transcript with the CRISPR locus from an unknown type I system.[8] The functional similarities between type I and II systems, such as the requirement of a PAM (protospacer adjacent motif) sequence, however, are inconsistent due to the lack of primary sequence similarity between Cas9 and other type I proteins.[8,34] Therefore, even in the lack of primary sequence similarity, Cas9 and type I proteins may overall conserve their secondary and tertiary structure, which together contribute toward their similar function.

This study aimed to understand the protein homology, sequence conservation, evolutionary relationships, and structure-function constraints among Cas3, Cas9, and Cas10, employing amino acid similarity search using blastp,[35] phylogenetic method using a maximum-likelihood method,[36] and estimating the ratios of the non-synonymous substitution rate (dN) and the synonymous substitution rate (dS) per site between the 2 homologous protein sequences using MATLAB.[37]

Materials and Methods

Bacterial genomes

The National Center for Biotechnology Information (NCBI) provides a collection of databases and bioinformatics tools for genome analysis. The nucleotide and protein databases were used from bacterial genomes which were completely sequenced and fully annotated. We examined 3 CRISPR-associated Cas protein families by identifying Cas protein homologs, phylogenetic analysis, and structure-function constraint analysis as summarized in the schematic diagram in Figure 1.

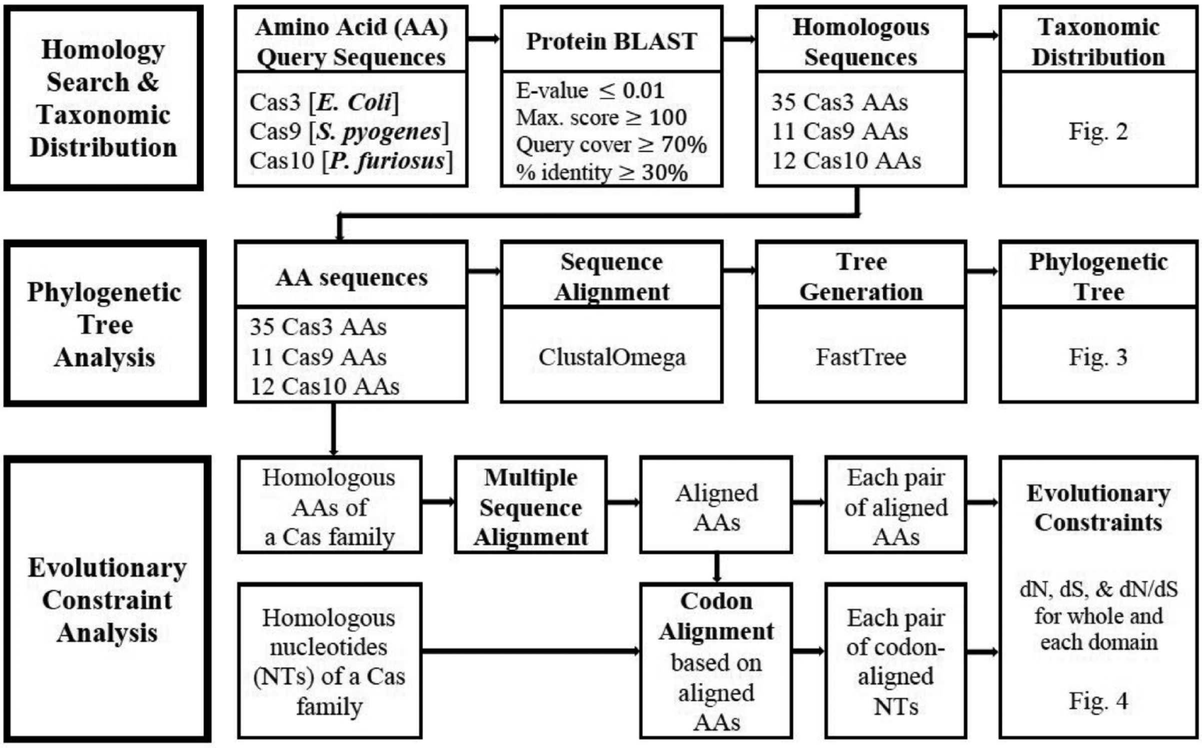

Figure 1. Workflow of homology search and taxonomic distribution, phylogenetic tree analysis, and evolutionary constraint analysis.

Identification of Cas protein homologs

The 3 Cas proteins, Cas3, Cas9, and Cas10, were examined in this study. Based on their common usage in previous studies of the CRISPR/Cas system, Cas3 of *Escherichia coli*,[38] Cas9 of *Streptococcus pyogenes*,[39] and Cas10 of *Pyrococcus furiosus*[40] were chosen as queries for DNA and non-redundant protein sequence similarity searches against the corresponding databases available at the NCBI using blastp.[35] The criteria for the homology search were as follows: E-value 10^{-2}, Maximum Score 100, Query Coverage 70%, and percent amino acid identity >30%. From the filtered homologous protein sequences, only 1 amino acid sequence from each bacterial species was downloaded. Species were classified into phyla, such as *Actinobacteria*, *Cyanobacteria*, *Firmicutes*, *Proteobacteria*, or *Spirochaetes*.

To perform both the phylogenetic analysis and the structure-function analysis, the identified homologs were further reduced based on the availability of the corresponding nucleotide sequences in the NCBI database. The numbers of resulting selected sequences for Cas3, Cas9, and Cas10 were 35, 11, and 13, respectively. All the 59 sequences are available on request.

Phylogenetic analysis

Phylogenetic analysis was performed using the Geneious Prime platform,[41] which consists of several different tools and plugins for molecular sequence analyses. First, a total of 58 proteins multiple protein sequences were directly downloaded to the Geneious and then aligned using a multiple-sequence alignment method, Clustal Omega[42] with a default setting in the Geneious platform. Clustal Omega is based on seeded guide trees and Hidden Markov model (HMM) profile-profile techniques to generate multiple alignments. Then, an unrooted phylogenetic tree was constructed using an approximate maximum-likelihood method, FastTree[43] with a default setting in the Geneious platform. Unlike usual approaches that store distance matrices, FastTree stores sequence profiles of tree's internal nodes and uses varied heuristics to quickly infer maximum-likelihood phylogenies for large number of sequences. The numbers above the branch points indicate the reliability of each split in the tree and these local support values estimated with the Shimodaira-Hasegawa test, in line with the SH-like local supports in PhyML3.0.[44] The tree is not drawn to scale; thus, branch lengths do not measure the number of substitutions per site.

Selective constraint analysis

For the constraint analysis for each of the 3 Cas families (Cas3, Cas9, and Cas10), a progressive multiple alignment (using multialgn function in MATLAB[37]) was applied, and both non-synonymous substitution rate (dN) and non-synonymous substitution rate (dS) between 2 homologous nucleotide sequences (using dnds function in MATLAB) were estimated. The ratio (dN/dS) of these mutation rates along the pair of sequences was used to predict the selective constraints between the 2 entire protein sequences in each system as well as each different protein domain within each system. The separation of different domains allows the evolutionary constraints separated by the regions of the protein that are critical for functionality. For each Cas family, evolutionary constraints were analyzed for the entire length of sequences as well as the sequences corresponding to specific domains. As the amino acid residue numbers of Cas3 domain borders of *Thermofibida fusca*[13] were available, the domain information of *T fusca* was used to align and map to the corresponding domains of the Cas3 protein of *E coli*. The coordinates of the Cas9 domain borders of *S pyogenes*[17,45] were used to identify the corresponding domain locations in 11 aligned Cas9 sequences. The coordinates of the Cas10 domain borders of *Pyrococcus furiosus*[20,21] were used to identify the corresponding domain location in 12 aligned Cas10 sequences.

Note that the total of $n(n-1)/2$ pairs can be made from n sequences in a Cas protein family. However, if the 2 sequences in a pair are too short, too divergent, or contain frame shifts, then saturation can be reached and the constraint values are unable to be estimated, resulting in Not-a-Number (NaN) values. Therefore, only valid constraint values estimated by *dnds* in MATLAB are used for the analysis. Each error bar summarizes evolutionary constraint values estimated for a specific domain in a Cas protein family, where the bar and the error represent the arithmetic mean and the standard deviation of the constraint values, respectively.

Results and Discussion

Distribution of Cas proteins across bacterial phyla

The percentage distribution of Cas3 protein in *Proteobacteria*, *Firmicutes Actinobacteria*, *Bacteroidetes* and other phyla are 29%, 20%, 9%, 9%, and 33%, respectively. The percentage distribution of Cas9 protein in *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, and other phyla are 33%, 26%, 3%, 12%, and 26%, respectively. The percentage distribution of Cas10 protein in *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, and other phyla are 23%, 20%, 3%, 11%, and 43%, respectively. Although the distribution of Cas3, Cas9, and Cas10 proteins among 3 phyla (*Proteobacteria*, *Firmicutes*, and *Actinobacteria*) are similar, significant differences exists among in the category of phylum *Bacteroidetes* and other phyla. The asymmetric distribution of Cas proteins in bacterial phyla is validated by a recent study where uneven distribution of CRISPR-Cas types was also reported.[6] The abundance of 3 Cas proteins in the 2 phyla (*Proteobacteria* and *Firmicutes*) is probably due to a higher number of sequenced genomes for these corresponding phyla present in the NCBI database. A previous study also discovered a similar bias in the number of sequenced genomes being from these 4 phyla.[46] The above findings demonstrate that Cas3, Cas9, and Cas10 proteins do not exhibit similar distributions among all major bacterial phyla; however, the distribution biases are not of the inherent characteristics of the respective phyla, instead the members of available complete genome sequences in the database (Figure 2).

Evolutionary relationships of Cas proteins

Across all 3 Cas protein families, both within and between the family members, there is a low to moderate amount of amino acid sequence conservation ranging from 30% to 50% amino acid identity. An unrooted phylogenetic tree as shown in Figure 3 exhibits the evolutionary relatedness of Cas-associated proteins of different bacterial species that possess those corresponding reference proteins.

Cas proteins were separated into 3 distinct clades, Cas3, Cas9, and Cas10, which revealed that these 3 proteins originated monophyletically. It was observed that the branch length of the clades representing Cas3 and Cas10 proteins is relatively longer than that of clades representing Cas9 proteins (results not shown in the tree), suggesting that the members of the Cas3 and Cas10 proteins evolved more rapidly than those of the members of Cas9 protein family. It is interesting to note that all clades representing Cas9 and Cas10 proteins diverged from a common clade of Cas3 proteins and later diverged into Cas9 and Cas10 protein families. The phylogenetic tree also reveals that Cas3 proteins represent the higher number of clades compared with the number of clades represented within Cas9 or Cas10 proteins. This finding is attributed maybe due to a large number of species representing Cas3 proteins included for the tree construction. Overall, the branch lengths and number of branches within the Cas3 family suggests that members of the Cas3 family evolved and diverged more rapidly than the members of the Cas9 or Cas10 protein families. Although, the phylogenetic tree of the Cas proteins was not reported earlier, a study suggests a common ancestry of the effector complexes of type I and type III Cas systems in which Cas3 and Cas10 belongs, respectively.[6]

Protein domains experience differential selection pressures

Neutral theory of molecular evolution accounts for the occurrence of both non-synonymous and synonymous substitutions.[47] The ratio between the 2 types of substitutions has been used to determine the strength of structure-function constraints.[48]

Specifically, the ratio (ie, dN/dS and denoted by ω) of the non-synonymous substitution rate (dN) to the synonymous substitution rate (dS) indicates the evolution of the structure-function constraint of the different domains or the whole protein. A dN/dS value is smaller than 1 means that the protein can accumulate more synonymous (silent) substitutions than non-synonymous (missense) substitutions. A dN/dS value approximately equal to 1 means that the mutations have no negative or positive effects on an organism's ability to survive

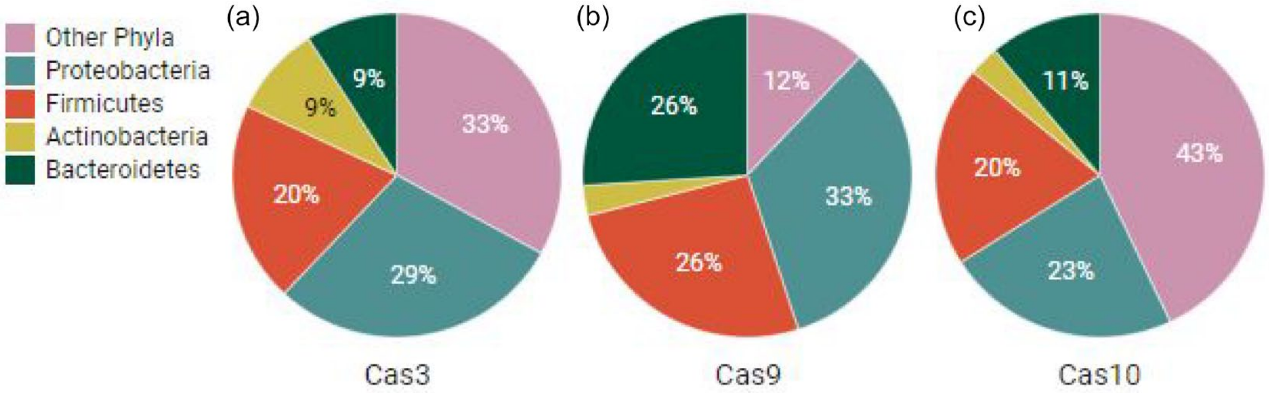

Figure 2. Distribution of (a) Cas3, (b) Cas9, and (c) Cas10 across bacterial phyla.

and reproduce.[47,48] In addition, neutral mutations alter amino acids without changing their chemical properties. Such neutral mutations can be fixed or fluctuate in a population. Finally, a dN/dS value greater than 1 indicates that the protein accumulates more missense mutations, which change amino acids with different chemical properties. Such mutations result in new functions under positive selection.

The rate of non-synonymous changes (dN), the rate of synonymous changes (dS), and the ratio of the non-synonymous to the synonymous changes (dN/dS) for the whole protein (corresponding to the first error bar in each subplot) and different protein domains (corresponding the remaining error bars in each subplot) for Cas3, Cas9, and Cas10 proteins are shown in Figure 4a to c, respectively. As shown in Figure 4a, Cas3 proteins have the 5 domains: an HD-nuclease domain, 2 Super Family 2 (SF2) helicase domains, a linker domain, and a C-terminal domain (CTD). Overall, Cas3 proteins experiences negative selection pressure, but different protein domains are functionally constrained differently. Of the 5 domains, the HD-nuclease domain and the 2 SF2 helicase domains (RECA1 and RECA2) appear to be experiencing negative selection, and they are highly conserved. This finding supports a previous observation that the helicase domain of Cas3 is highly conserved among the members of the Cas3 protein family.[49] The SF2 helicase domain is an ATPase responsible for unwinding the target DNA strand. This ATPase activity allows the HD-nuclease domain to cleave the ssDNA strand of the target sequence.[13] The importance of these domains for the destruction of foreign DNA makes it obvious that these domains are under negative selective pressure. Conversely, the CTD shows a dN/dS value above 1, which indicates the positive selection pressure in that region and the low conservation. The CTD appears to be responsible for interacting with Csel, which is a protein in the CASCADE array that functions to recognize the PAM sequence in target DNA.[50] The rapid evolution of this domain has likely occurred to accommodate the high degree of variation in PAM sequence recognition across bacterial species.

As shown in Figure 4b, the structure-function constraint analysis for Cas9 proteins has similar conservation patterns to those seen for Cas3 proteins. Overall, Cas9 proteins experience negative selection. It has 9 different domains, where 7 domains (including 2 REC1/RECI, bridge helix (BH), HNH, and 3 RuvCs) maintain high levels of sequence conservation, experience negative selective pressure with dN/dS less than 0.5, and correspond to the functions such as binding, cleaving, and destructing the target DNA.[51] As shown in Figure 4, whole protein sequences in each Cas family are highly constrained; thus, they are experiencing negative selection, where Cas3 and Cas9 are relatively more constrained than Cas10. All these domains are crucial to the functionality of the whole system, so a higher level of constraint ($dN/dS < 1$) is expected. The PAM-interacting (PI) domain of Cas9 appears to be experiencing negative selection. As previously discussed for Cas3, a higher level of mutations is expected to allow the level of variations in the PAM sequences; however, Cas9 PAM seems different and shows less sequence variation.

The REC1/RECI domain which remains uncharacterized also demonstrated a dN/dS value above 1, indicating that it has experienced positive selective pressure. A similar finding, stating that the REC lobe is one of the least conserved regions across all members of the Cas9 family, was also reported in a previous study.[51] This further suggests that a high level of conservation of the REC2/RECII domain might not be required for the functionality of the system.

As shown in Figure 4c, Cas10 proteins have different 5 domains including HD nuclease, Fingers, Zn Finger/motif, Palm/RRM and Thumb. Although Cas10 proteins overall experience negative selection as seen in the case of Cas3 and Cas9 proteins, their protein domains show evolutionary constraints with dN/dS values close to 1 or above 1, which are relatively higher number than dN/dS values for most domains in both Cas3 and Cas9. It indicates that Cas10 domains experience a lower level of sequence conservation as well as less structure-function constraint. A previous study similarly demonstrated that 85% of the Cas10 protein family have

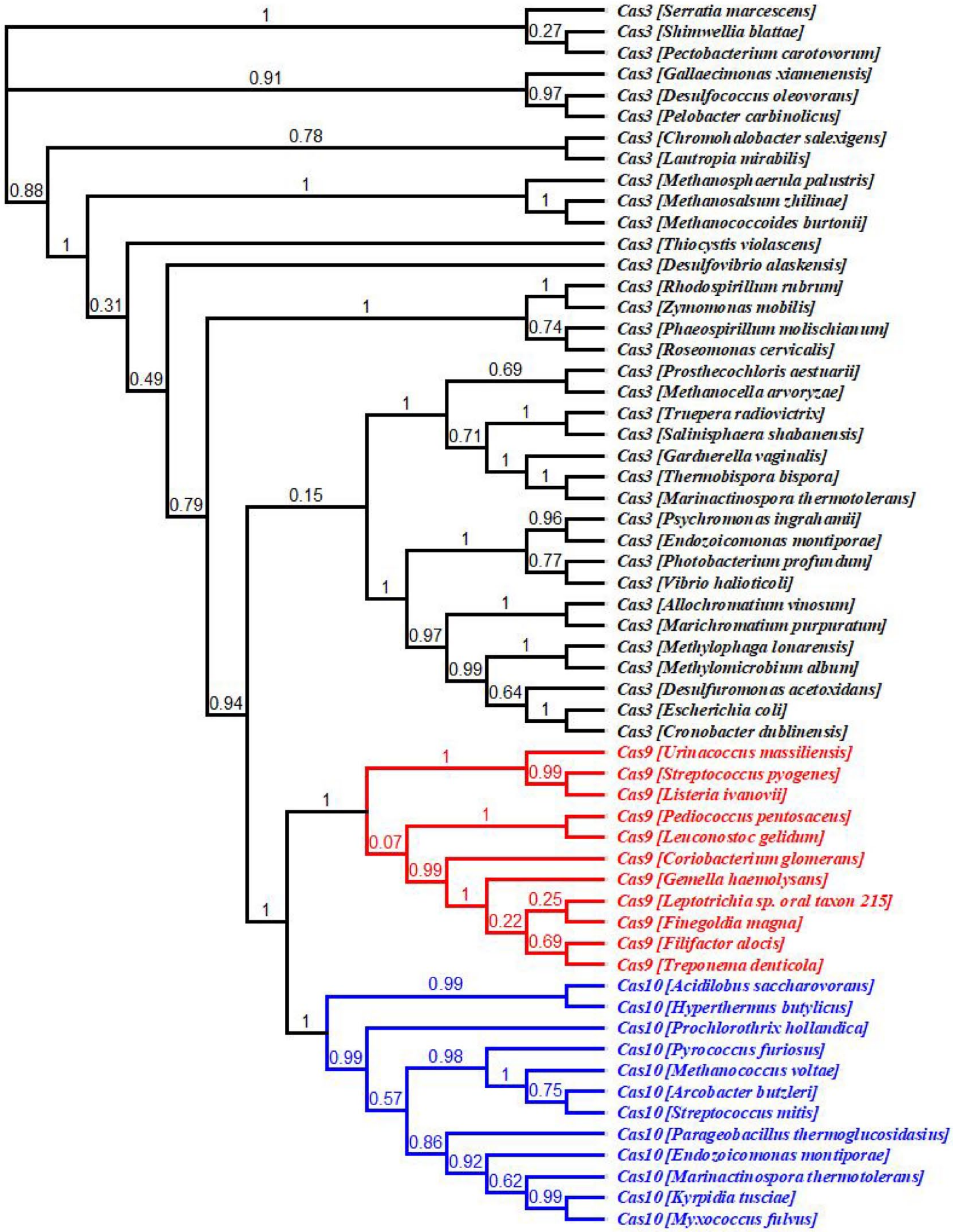
Cas3 [Serratia marcescens]
Cas3 [Shimwellia blattae]
Cas3 [Pectobacterium carotovorum]
Cas3 [Gallaecimonas xiamenensis]
Cas3 [Desulfococcus oleovorans]
Cas3 [Pelobacter carbinolicus]
Cas3 [Chromohalobacter salexigens]
Cas3 [Lautropia mirabilis]
Cas3 [Methanosphaerula palustris]
Cas3 [Methanosalsum zhilinae]
Cas3 [Methanococcoides burtonii]
Cas3 [Thiocystis violascens]
Cas3 [Desulfovibrio alaskensis]
Cas3 [Rhodospirillum rubrum]
Cas3 [Zymomonas mobilis]
Cas3 [Phaeospirillum molischianum]
Cas3 [Roseomonas cervicalis]
Cas3 [Prosthecochloris aestuarii]
Cas3 [Methanocella arvoryzae]
Cas3 [Truepera radiovictrix]
Cas3 [Salinisphaera shabanensis]
Cas3 [Gardnerella vaginalis]
Cas3 [Thermobispora bispora]
Cas3 [Marinactinospora thermotolerans]
Cas3 [Psychromonas ingrahamii]
Cas3 [Endozoicomonas montiporae]
Cas3 [Photobacterium profundum]
Cas3 [Vibrio halioticoli]
Cas3 [Allochromatium vinosum]
Cas3 [Marichromatium purpuratum]
Cas3 [Methylophaga lonarensis]
Cas3 [Methylomicrobium album]
Cas3 [Desulfuromonas acetoxidans]
Cas3 [Escherichia coli]
Cas3 [Cronobacter dublinensis]
Cas9 [Urinacoccus massiliensis]
Cas9 [Streptococcus pyogenes]
Cas9 [Listeria ivanovii]
Cas9 [Pediococcus pentosaceus]
Cas9 [Leuconostoc gelidum]
Cas9 [Coriobacterium glomerans]
Cas9 [Gemella haemolysans]
Cas9 [Leptotrichia sp. oral taxon 215]
Cas9 [Finegoldia magna]
Cas9 [Filifactor alocis]
Cas9 [Treponema denticola]
Cas10 [Acidilobus saccharovorans]
Cas10 [Hyperthermus butylicus]
Cas10 [Prochlorothrix hollandica]
Cas10 [Pyrococcus furiosus]
Cas10 [Methanococcus voltae]
Cas10 [Arcobacter butzleri]
Cas10 [Streptococcus mitis]
Cas10 [Parageobacillus thermoglucosidasius]
Cas10 [Endozoicomonas montiporae]
Cas10 [Marinactinospora thermotolerans]
Cas10 [Kyrpidia tusciae]
Cas10 [Myxococcus fulvus]
1
0.27
0.91
0.97
0.78
0.88
1
1
1
0.31
1
1
0.74
0.49
0.69
1
1
0.71
1
1
0.15
0.79
0.96
1
0.77
1
1
0.97
1
0.99
0.64
1
0.94
1
0.99
1
1
0.07
0.99
1
0.25
0.22
0.69
1
0.99
1
0.99
0.98
1
0.75
0.57
0.86
0.92
0.62
0.99

Figure 3. The phylogenetic tree of Cas3, Cas9, and Cas10 proteins.

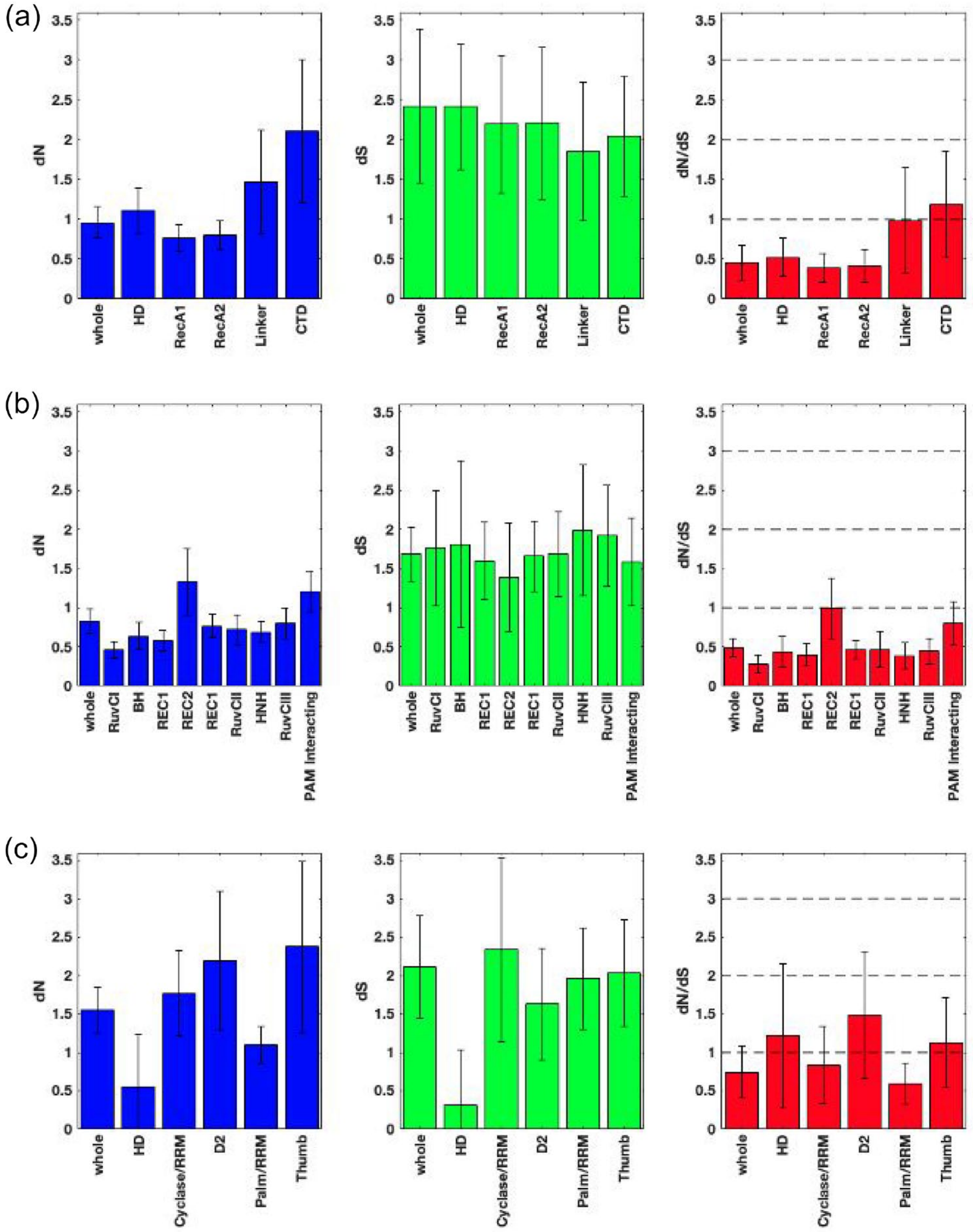

Figure 4. The rate of non-synonymous substitutions (dN), the rate of synonymous substitutions (dS), and the dN/dS ratio for (a) Cas3, (b) Cas9, and (c) Cas10. The structure of *Thermofibida fusca* (*T fusca*) Cas3 consists of 5 domains, including HD (residues 1-258), RecA1 (residues 259-545), RecA2 (residues 546-777), Linker (residues 778-834), and CTD (residues 834-944). Through a pair-wise alignment with *T fusca*, the domain positions of *Escherichia coli* (*E coli*) Cas3 are obtained as follows: HD (residues 1-267), RecA1 (residues 259-545), RecA2 (residues 546-777), linker (residues 778-834), and CTD (residues 834-944).

conserved polymerase active-site motifs; however, only 36% of the Cas10 protein family have conserved HD-nuclease domain.[52] Therefore, it can be suggested that the Cas10 protein is experiencing positive selection, and its various domains evolved more rapidly than the domains of Cas3 and Cas9 protein families. As the function of the Cas10 protein remains poorly understood, it is difficult to correlate their functions with differences of structure-function constraints. This further corroborates the phylogenetic relationship that these proteins have evolved rapidly and show little similarity in functionality across protein families.

Streptococcus pyogenes Cas9 is organized with recognition (REC) and nuclease (NUC) lobes. The REC lobe consists of a long alpha-helix called BH and multiple alpha-helical REC domains (REC1 [residues 94-179], REC2 [residues 180-307], REC1 [residues 308-717]), whereas the nuclease (NUC) domain comprises 2 endonuclease domains (HNH [residues 775-908]) and 3 RuvCs (RuvCI [residues 1-59], RuvCII [residues 718-774], and RuvCIII [residues 909-1098]) and a PI domain (residues 1099-1368).

Pyrococcus furiosus Cmr2 (PDF ID 4W8Y) Cas10 comprises HD domain (residues 1-216), Cyclase/RRM (residues 216-503) including Zn finger domain at the end, D2 (residues 503-593) whose residues are based on Zhu and Ye[20] and Manav et al[21] but whose function is unclear, Palm/RRM (residues 593-763) domain, and Thumb (residues 764-871) domain.

Conclusion

Based on the results of protein conservation, phylogenetic tree, and structure-function analyses, the following conclusions are made: CRISPR-associated Cas3, Cas9, and Cas10 proteins are prevalent in bacteria and archaea, and they are distributed among major bacterial phyla. All 3 proteins have their monophyletic origins with Cas9 and Casl0 first being commonly diverged from Cas3, and the 2 were later separated into separate protein families. The analysis of the branch lengths and the number of clades suggest that members of the Cas3 family evolved more rapidly than the members of the Cas9 or Cas10 protein families. Cas3 and Cas9 proteins, including their most of functionally interactive domains, experience negative selection, whereas Cas10 protein along with its all functionally interactive domains experience positive selection. The result supports that Cas10 protein and its functionally interactive domains may have evolved new molecular functions that yet to be characterized.

Future work includes similar analyses of other proteins associated with type I, II and III systems that would provide further information about the origins and functions of the CRISPR systems. Also, structure-function constraints of the different functionally interactive domains may be further exploring the scope of an alternate and improved genome editing system to the currently used CRISPR-Cas9 system. In addition, the utilization of the CRISPR-Cas3/Cas10 may provide additional clues to the recently discovered immunity to *S pyogenes* Cas9 and *Staphylococcus aureus* Cas9 systems.[53]

Acknowledgements

The authors acknowledge the contribution of Dorcie Gillette for conducting the data analysis, and the financial support for her thesis project from the Honors College, Sam Houston State University.

Author Contributions

Weerakkody Ranasinghe contributed for the manuscript preparation and revision. Dorcie Gillette contributed for the initial data collection and analysis. Alexis Ho contributed for the manuscript preparation and proof reading. Hyuk Cho contributed for the study design, data analysis, manuscript preparation and revision. Madhusudan Choudhary contributed for the study conception, study design, manuscript preparation and revision.

ORCID iD

Madhusudan Choudhary https://orcid.org/0000-0001-6217-7491

REFERENCES

1. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. *Science*. 2014;346:1258096. doi:10.1126/science.1258096
2. Li GL, Zhong CL, Ni S, et al. Establishment of porcine Xist knockout model using CRISPR/Cas9 system. *Yi Chuan*. 2016;38:1081-1089. doi:10.16288/j.yczz.16-137
3. Wang Y, Zhang ZT, Seo SO, et al. Bacterial genome editing with CRISPR-Cas9: deletion, integration, single nucleotide modification, and desirable "clean" mutant selection in Clostridium beijerinckii as an example. *ACS Synth Biol*. 2016;5:721-732. doi:10.1021/acssynbio.6b00060
4. Mojica FJ, Juez G, Rodriguez-Valera F. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified Pst I sites. *Mol Microbiol*. 1993;9:613-621. doi:10.1111/j.1365-2958.1993.tb01721.x
5. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337:816-821. doi:10.1126/science.1225829
6. Makarova KS, Wolf YI, Iranzo J, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*. 2020;18:67-83. doi:10.1038/s41579-019-0299-x
7. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007;315:1709-1712. doi:10.1126/science.1138140
8. Jansen R, van Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*. 2002;43:1565-1575. doi:10.1046/j.1365-2958.2002.02839.x
9. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*. 2017;37:67-78. doi:10.1016/j.mib.2017.05.008
10. Jiang F, Doudna JA. CRISPR–Cas9 structures and mechanisms. *Annu Rev Biophys*. 2017;46:505-529. doi:10.1146/annurev-biophys-062215-010822
11. Rath D, Amlinger L, Rath A, Lundgren M. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*. 2015;117:119-128. doi:10.1016/j.biochi.2015.03.025
12. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system: Cas3 nuclease/helicase. *EMBO J*. 2011;30:1335-1342. doi:10.1038/emboj.2011.41
13. Huo Y, Nam KH, Ding F, et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol*. 2014;21:771-777. doi:10.1038/nsmb.2875
14. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. 2011;471:602-607. doi:10.1038/nature09886
15. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A*. 2012;109. doi:10.1073/pnas.1208507109
16. Mao Y, Zhang H, Xu N, Zhang B, Gou F, Zhu JK. Application of the CRISPR-Cas system for efficient genome engineering in plants. *Mol Plant*. 2013;6:2008-2011. doi:10.1093/mp/sst121
17. Allemailem KS, Alsahli MA, Almatroudi A, et al. Current updates of CRISPR/Cas9-mediated genome editing and targeting within tumor cells: an innovative strategy of cancer management. *Cancer Commun*. 2022;42:1257-1287. doi:10.1002/cac2.12366
18. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct*. 2011;6:38. doi:10.1186/1745-6150-6-38

19. Rouillon C, Zhou M, Zhang J, et al. Structure of the CRISPR interference complex CSM reveals key similarities with Cascade. *Mol Cell*. 2013;52:124-134. doi:10.1016/j.molcel.2013.08.020
20. Zhu X, Ye K. Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. *FEBS Lett*. 2012;586:939-945. doi:10.1016/j.febslet.2012.02.036
21. Manav MC, Van LB, Lin J, Fuglsang A, Peng X, Brodersen DE. Structural basis for inhibition of an archaeal CRISPR–Cas type I-D large subunit by an anti-CRISPR protein. *Nat Commun*. 2020;11:5993. doi:10.1038/s41467-020-19847-x
22. Sinkunas T, Gasiunas G, Siksnys V. Cas3 nuclease–helicase activity assays. *Methods Mol Biol*. 2015;1311:277-291. doi:10.1007/978-1-4939-2687-9_18
23. Redding S, Sternberg SH, Marshall M, et al. Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. *Cell*. 2015;163:854-865. doi:10.1016/j.cell.2015.10.003
24. Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*. 2005;1:e60. doi:10.1371/journal.pcbi.0010060
25. Gilbert LA, Larson MH, Morsut L, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*. 2013;154:442-451. doi:10.1016/j.cell.2013.06.044
26. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc*. 2013;8:2180-2196. doi:10.1038/nprot.2013.132
27. Jinek M, Jiang F, Taylor DW, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*. 2014;343:1247997. doi:10.1126/science.1247997
28. Makarova KS. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*. 2002;30:482-496. doi:10.1093/nar/30.2.482
29. Liu TY, Liu JJ, Aditham AJ, Nogales E, Doudna JA. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. *Nat Commun*. 2019;10:3001. doi:10.1038/s41467-019-10780-2
30. Elmore JR, Sheppard NF, Ramia N, et al. Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR–Cas system. *Genes Dev*. 2016;30:447-459. doi:10.1101/gad.272153.115
31. Estrella MA, Kuo F-T, Bailey S. RNA-activated DNA cleavage by the Type III-B CRISPR–Cas effector complex. *Genes Dev*. 2016;30:460-470. doi:10.1101/gad.273722.115
32. Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas Siksnys ČV. Spatiotemporal control of Type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition. *Mol Cell*. 2016;62:295-306. doi:10.1016/j.molcel.2016.03.024
33. Makarova KS, Wolf YI, Alkhnbashi OS, et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat Rev Microbiol*. 2015;13:722-736. doi:10.1038/nrmicro3569
34. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. *Phil Trans R Soc B*. 2019;374:20180087. doi:10.1098/rstb.2018.0087
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-410. doi:10.1016/S0022-2836(05)80360-2
36. Lin Y, Hu F, Tang J, Moret BM. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. *Pac Symp Biocomput*. 2013:285-296.
37. The MathWorks Inc., Natick, MA. Accessed August 23, 2024. https://www.mathworks.com
38. Majsec K, Bolt EL, Ivančić-Baće I. Cas3 is a limiting factor for CRISPR-Cas immunity in Escherichia coli cells lacking H-NS. *BMC Microbiol*. 2016;16:28. doi:10.1186/s12866-016-0643-5
39. Nozawa T, Furukawa N, Aikawa C, et al. CRISPR inhibition of prophage acquisition in Streptococcus pyogenes. *PLoS ONE*. 2011;6:e19543. doi:10.1371/journal.pone.0019543
40. Shao Y, Cocozaki AI, Ramia NF, Terns RM, Terns MP, Li H. Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex. *Structure*. 2013;21:376-384. doi:10.1016/j.str.2013.01.002
41. Kearse M, Moir R, Wilson A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647-1649. doi:10.1093/bioinformatics/bts199
42. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539. doi:10.1038/msb.2011.75
43. Price MN, Dehal PS, Arkin AP. FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 2010;5:e9490. doi:10.1371/journal.pone.0009490
44. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59:307-321. doi:10.1093/sysbio/syq010
45. Antao AM, Karapurkar JK, Lee DR, Kim KS, Ramakrishna S. Disease modeling and stem cell immunoengineering in regenerative medicine using CRISPR/Cas9 systems. *Comput Struct Biotechnol J*. 2020;18:3649-3665. doi:10.1016/j.csbj.2020.11.026
46. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. *Genome Biol*. 2002;3:REVIEWS0003. doi:10.1186/gb-2002-3-2-reviews0003
47. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 1977;267:275-276. doi:10.1038/267275a0
48. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*. 2000;15:496-503. doi:10.1016/S0169-5347(00)01994-7
49. Loeff L, Brouns SJ, Joo C. Repetitive DNA reeling by the Cascade-Cas3 complex in nucleotide unwinding steps. *Mol Cell*. 2018;70:385-394e3. doi:10.1016/j.molcel.2018.03.031
50. Gong B, Shin M, Sun J, et al. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A*. 2014;111:16359-16364. doi:10.1073/pnas.1410806111
51. Nishimasu H, Ran FA, Hsu PD, et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*. 2014;156:935-949. doi:10.1016/j.cell.2014.02.001
52. Wiegand T, Wilkinson R, Santiago-Frangos A, Lynes M, Hatzenpichler R, Wiedenheft B. Functional and phylogenetic diversity of Cas10 proteins. *CRISPR J*. 2023;6:152-162. doi:10.1089/crispr.2022.0085
53. Charlesworth CT, Deshpande PS, Dever DP, et al. Identification of preexisting adaptive immunity to Cas9 proteins in humans. *Nat Med*. 2019;25:249-254. doi:10.1038/s41591-018-0326-x