

BIOINFORMATICS AND PHYLOGENETIC
ANALYSIS OF
CELLULAR DEFENSE SYSTEMS

by
Murat Buyukyoruk

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Microbiology and Cell Biology

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2023

©COPYRIGHT

by

Murat Buyukyoruk

2023

All Rights Reserved

ACKNOWLEDGEMENTS

My doctoral work has been an adventurous journey that has only been possible thanks to the help and support of many people. First and foremost, I'd like to thank my advisor, Blake Wiedenheft. Blake has taught me how to ask good scientific questions, to consider why one would care for the answers, and how one can convey them precisely and clearly to others. His enthusiasm for scientific discovery is infectious, and some of my favorite times in the Wiedenheft lab were spent brainstorming during lab meetings or impromptu meetings in Blake's office.

I would like to express my heartfelt gratitude to my committee members; Dr. Michael Franklin, Dr. Edward Schmidt, and Dr. Mark Young, for their unwavering support, invaluable efforts, and profound scientific insights.

I would also like to thank the colleagues who helped me to my successes and commiserated with me following failures. MC, Royce, Sarah, Heini, Paul, Andrew, Artem and Anna all made the Wiedenheft lab an exciting place to work and were eternally patient mentors as I learned how to do microbiological research. My graduate student peers and the undergraduates who worked in our lab were likewise instrumental in my development as a scientist.

Finally, I want to acknowledge those whose love and support were encouraging when experiments failed. My friends, Reha, Berrak, Batu and Galip, kept me sane and grounded. A special thanks to my parents, Atiye and Hayri, my cousins, Gamze and Can, and my aunt and uncle Aysun, and Saffet. Finally, special thanks to my grandmother Tevhide for all her support and good wishes. Family has always been a cornerstone in my life, and I couldn't have completed this work without you.

TABLE OF CONTENTS

INTRODUCTION	1
RNA Guides: Achieving Target Recognition Through CRISPR and Argonautes	1
RNA-based Recognition by miRNAs and siRNAs in Argonautes	3
RNA-based Recognition by crRNAs in CRISPR-Cas Systems	7
Summary	10
References	12
TYPE I-F CRISPR-CAS PROVIDES PROTECTION FROM DNA, BUT NOT RNA PHAGES	14
Contribution of Authors and Co-Authors	14
Manuscript Information	15
Manuscript	17
Methods and Materials	20
Strains, Plasmids and Growth Conditions	20
Synthetic CRISPR Array Construction	21
Phage Replication and Plaque Assays	21
Spot Plaque Assay	22
Quantification and Statistical Analysis	22
Acknowledgements	22
References	23
Supplemental Figures	24
DISTRIBUTION AND PHASING OF SEQUENCE MOTIFS THAT FACILITATE CRISPR ADAPTATION	26
Contribution of Authors and Co-Authors	26
Manuscript Information	27
Abstract	29
Introduction	30
Results	33
Identification and Distribution of IHF and Upstream Motifs	33
IHF-binding Sites and Upstream Motif are Critical for Efficient Integration	38
Tn7-associated I-F3 Leaders Contain Motifs Needed for Efficient Integration	41
Phased Leader Motifs Facilitate CRISPR Adaptation	42
Discussion	45
Acknowledgements	48
References	49

TABLE OF CONTENTS CONTINUED

Supplemental Figures.....	57
Materials and Methods.....	63
Bacterial Strains	63
Modelling the I-F CRISPR Integration Intermediate Complex	63
Building a Database of CRISPR Leaders	64
Identification of Conserved DNA Motifs in CRISPR Leaders	64
Phylogenetic Analyses	65
Plasmid Construction	66
IHF Expression and Purification.....	67
Cas1-2/3 Expression and Purification.....	68
Nucleic Acid Preparation	69
<i>In vitro</i> Integration Assays	70
High-throughput Sequencing of <i>In vitro</i> Integration Products	71
Quantification and Statistical Analysis	72
IDENTIFICATION OF SUBTYPE- AND TAXON-SPECIFIC CRISPR LEADER MOTIFS ENHANCES CRISPR DETECTION AND ANNOTATION.....	73
Contribution of Authors and Co-Authors	73
Manuscript Information	74
Abstract	76
Introduction.....	77
Results.....	80
Identification and Taxonomic Distribution of 77 CRISPR Leader Motifs.....	80
Leader Motifs Complement Repeat- and Cas-based Annotations of CRISPRs	81
CRISPR Leader Motifs Enable Annotation of Short and Degenerate Putative CRISPR Arrays.....	84
Discussion	87
Methods.....	90
Building a Database of CRISPR Leaders	90
Subtype Annotation of CRISPR Arrays in CRISPR Leader Database.....	90
Identification of Host-taxonomy.....	91
Identification of Conserved DNA Motifs in CRISPR Leaders	91
Scoring Matrix for Leader-motif-assisted CRISPR Annotation	92
16S Ribosomal RNA Analysis of Bacterial and Archaeal Genomes with CRISPR.....	94
Identification of Short and Degenerate Arrays	94
Acknowledgements.....	95

TABLE OF CONTENTS CONTINUED

References.....	96
Supplemental Tables and Figures	100
 CLARIFYING CRISPR: WHY REPEATS IDENTIFIED IN THE HUMAN GENOME SHOULD NOT BE CONSIDERED CRISPRS.....	 122
Contribution of Authors and Co-Authors	122
Manuscript Information	123
Abstract	125
Introduction.....	126
Methods.....	127
CRISPR Predictions.....	127
Identification of <i>cas</i> Genes in the Human Genome	128
Comparison of Repeat Length Distribution of CRISPRs from Prokaryotes and hCRISPRs	129
Results and Discussion	129
Acknowledgements.....	135
References.....	136
Supplemental Information	138
 PHYLOGENETIC AND FUNCTIONAL ANALYSIS OF PROKARYOTIC ARGONAUTE PROTEINS WITH DIVERSE FUNCTIONS.....	 142
Contribution of Authors and Co-Authors	142
Manuscript Information	143
Abstract	145
Introduction.....	146
Results and Discussion	148
Expansion of Ago-associated Effector Domains	148
N-terminal of eAgos Usually Contains Unstructured Regions with NLS and NES Signals.....	152
APAZ: Absence of PAZ	153
Predicted Modular Mechanisms of Short pAgos in Response to Cellular Stress	156
Conclusion	158
Acknowledgement	159
Methods.....	159
Building a Database of Ago Proteins	159
Identification of APAZ Genes in the Genomes Containing Agos.....	160
Ago and APAZ Phylogenies	160
Domain Annotation of Argonaute and APAZ Sequences	160
References.....	162

TABLE OF CONTENTS CONTINUED

Supplemental Figures.....	166
GENERAL CONCLUSION AND OUTLOOK	167
From CRISPR Interference to Regulation of CRISPRs	167
Looking at Genomes Outside of the Prokaryotes to Find Evidence of CRISPRs	169
Diversity of Prokaryotic Argonautes	170
Concluding Remarks.....	171
CUMULATIVE REFERENCES CITED.....	172

LIST OF TABLES

Table	Page
Table 3.1. Subtype distribution of microbial CRISPRs and identified IHF binding sites.	34
Table 3.2. Cas1 residues that interact with canonical I-E UM are only conserved in Cas1s associated with IHF-regulated I-E CRISPRs.	38
Table S4.1. Subtype distribution of microbial CRISPRs.	100
Table S4.2. Distribution of CRISPR leader sequences according to subtype and host taxonomy.	101
Table S4.3. Subtype distribution of CRISPR arrays predicted from the "Relaxed" search with CRISPRDetect and annotation with repeat-, <i>cas</i> - and leader-based methods.	102
Table S4.4. Detected Leader motifs with previously known functions.	121

LIST OF FIGURES

Figure	Page
Figure 1.1. The biogenesis, loading and targeting of RNA in distinct RNA-based systems.....	2
Figure 1.2. Presentation of seed sequences in guide strands and target cleavage by Ago.....	5
Figure 1.3. The key steps of CRISPR-Cas immunity.	8
Figure 1.4. Presentation of target binding in guide strands, conformational change of surveillance complex (Csy complex) and target cleavage by recruited Cas2/3 trans-acting nuclease/helicase.....	10
Figure 2.1. Testing the efficiency of type I-F CRISPR defense against RNA and DNA phages.....	19
Figure S2.1. CRISPR design.....	25
Figure 3.1. Spacing of sequence motifs in CRISPR leaders has mechanistic implications for the mechanism of spacer integration.	32
Figure 3.2. IHF-directed CRISPR adaptation is widespread.	36
Figure 3.3. IHF binding sites and upstream motif are critical for efficient integration.	39
Figure 3.4. Phase-dependent CRISPR adaptation.....	43
Figure S3.1. Naturally occurring indels preserve the phase of leader motifs, Related to Figures 3.2-3.4 and Table 3.1.	57
Figure S3.2. DNA-interacting residues are conserved between <i>E. coli</i> and <i>P. aeruginosa</i> IHF proteins. Related to Figure 3.3.	58
Figure S3.3. High-throughput sequencing and quantification of prespacer integration events. Related to Figures 3.3 and 3.4.....	59
Figure S3.4. IHF binds I-F leaders to recruit an upstream motif needed for efficient prespacer integration. Related to Figure 3.3.....	60
Figure S3.5. Summary of differential evolution rates between co-existing CRISPR loci. Related to Figure 3.3.....	61

LIST OF FIGURES CONTINUED

Figure	Page
Figure S3.6. Leader motifs involved in CRISPR evolution are phase sensitive. Related to Figure 3.4.....	62
Figure 4.1. Subtype and taxonomy-specific leader motifs are widespread in bacteria and archaea.....	79
Figure 4.2. Leader motifs complement repeat- and Cas-based annotation of CRISPRs.	82
Figure 4.3. Leader-based annotation improves the classification of short and degenerate CRISPR arrays.....	85
Figure 4.4. Distribution of array, repeat and spacer lengths predicted by an edited version of CRISPRDetect coupled with leader-based annotation.....	86
Figure S4.1. Each predicted motif possesses narrow distribution in the group CRISPR leader sequences that they originally identified.	103
Figure S4.2. PWMs predicted from de novo motif prediction analysis that represents different occurrences of IHF binding site in different subtypes.	114
Figure S4.3. Comparison of the motif-assisted array predictions against the high-confidence predictions generated by traditional classification methods.	115
Figure S4.4. Leader Motif-assisted Annotation in Non-leader Sequences yields less than 0.6% false-annotation.....	117
Figure S4.5. CRISPR Subtype-specific distribution of array, repeat and spacer lengths predicted from edited version of CRISPRDetect coupled with leader motif-assisted annotation.	119
Figure 5.1. Comparison of architectural features in CRISPRs and hCRISPRs.....	126
Figure 5.2. Repetitive elements in the human genome do not contain architectural features consistent with CRISPR loci.	131
Figure S5.1. Previously reported CRISPR-like elements.	141
Figure 6.1. Comparative Analysis of Argonaute Protein Architectures.....	150

LIST OF FIGURES CONTINUED

Figure	Page
Figure 6.2. Structural variations in prokaryotic argonautes highlighting the presence or absence of the PAZ domain.	154
Figure 6.3. Proposed model of modular recruitment of APAZ-effectors by short pAgos.	157
Figure S6.1. Illustration of the relationship between short pAgos and multiple APAZ effectors.....	166

ABSTRACT

Prokaryotes are under constant threat from foreign genetic elements such as viruses and plasmids. To defend themselves against these genetic invaders prokaryotes have evolved extensive defense mechanisms, including CRISPR-Cas systems and prokaryotic Argonautes. Both systems utilize guide nucleic acids to precisely target complementary foreign nucleic acids, showcasing the versatility of nucleic acid-guided genetic regulation. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) loci are repeat sequences that contain fragments of previously encountered genetic parasites. These loci are frequently flanked by CRISPR-associated (*cas*) genes essential for providing immunity against these invaders. In addition to the *cas* genes, CRISPRs often feature leader sequences positioned upstream of CRISPR loci, playing a crucial role in transcription and adaptation. While the diversity of CRISPR loci and *cas* genes has been extensively investigated, the diversity of CRISPR leaders has not been evaluated with the same level of rigor. Comprehensive bioinformatic evaluations have depicted diverse CRISPR-Cas system types throughout bacterial and archaeal genomes, highlighting the dynamic evolution between phages and bacteria. Although the CRISPR-Cas system's versatility has been co-opted for gene editing and diagnostics, current identification methods are limited. This study focusses on the diversity of CRISPR leader sequences and how the conserved sequence motifs can be used in enhancing detection capabilities, presenting potential biotechnological innovations. Similar to CRISPRs, Argonaute proteins rely on guide nucleic acids to identify complementary target sequences. Prokaryotic Argonautes (pAgos) are phylogenetically diverse which have often been overlooked. This research uncovers the rich diversity of pAgos, particularly focusing on their association with various genes and their associated effector proteins. This intricate relationship suggests a potentially multifaceted role for some pAgos in prokaryotic defense mechanisms.

PREFACE AND OUTLINE

Bacteriophages are the most abundant biological entities on earth. Due to the abundance and the diversity of phages, bacteria are under constant threat of infection. It has been estimated that approximately 10^{23} bacterial cell gets infected every second. To defend themselves against these threats, bacteria have evolved an arsenal of defense mechanisms that inhibit the phage life cycle.

Prokaryotic defense systems can be divided into three general strategies: (i) preventing cell entry that can be accomplished through mechanisms such as by alteration of entry receptor or prevention of viral genome injection, (ii) abortive infection that can be achieved by either altruistic suicide of the cell or leading to the dormancy of infected cells to protect the rest of the colony, and (iii) cellular immunity that targets phage nucleic acids, which includes both innate (i.e., Restriction-Modification systems, Argonautes) and adaptive immunity (CRISPR-Cas). Restriction-modification (R-M) and abortive infection (Abi) systems were the focus of early bacterial defense research until the discovery of CRISPR-Cas systems, which revealed adaptive immunity in prokaryotes akin to that in eukaryotes. However, CRISPR-Cas systems appear to be unique to prokaryotes, marking a significant departure from the simpler defense mechanisms previously understood.

CRISPR-Cas systems are unique in their ability to provide inheritable and adaptable defense against invaders. CRISPR arrays consist of short spacer fragments of foreign DNA inserted between palindromic repeats. One side of the CRISPR loci contains an A-T rich region called CRISPR leader. Next to this locus lie the CRISPR-associated (*cas*) genes that code for the protein machinery of the CRISPR systems. CRISPR-Cas systems provide immunity through three

universally conserved stages. Adaptation occurs during a phage infection when short segments of foreign DNA are integrated as new spacers into the CRISPR locus. This unique storage step is orchestrated by the nearly universal Cas1 and Cas2 proteins. Intriguingly, Cas1-2 integrase does not randomly select any repeat within the loci for the site of the integration of a protospacer. Instead, Cas1-2 integrates the DNA at the first repeat, creating a chronological archive of past phage infections. During the second stage, the CRISPR locus is transcribed into a single RNA strand that is processed by Cas proteins into individual crRNAs (CRISPR RNAs). Sequence motifs in the CRISPR leader function as genetic landmark to regulate integration at the first CRISPR repeat, transcription of the CRISPR locus, and processing of the first crRNA. In the final stage of CRISPR immunity, the surveillance complexes carrying a guide crRNA identify targets that are complementary to the guide. Target recognition activates intrinsic nuclease activity in the surveillance complex or recruits a trans-acting nuclease.

Large scale bioinformatic analyses have identified thousands of CRISPR-Cas systems in the sequences of bacterial and archaeal genomes, which can be broadly classified into two classes, six types, and over 33 subtypes. Each CRISPR system type is characterized by unique biochemical activities, highlighting the evolutionary arms race between phages and bacteria. Over the past decade the wide array of biochemical activities of CRISPR-Cas systems has been repurposed for revolutionary applications in gene editing, diagnostics, and DNA and RNA recording devices. This surge in applications has spurred a demand for enhanced CRISPR detection tools to uncover the natural diversity of bacterial and archaeal adaptive immune systems, paving the way for innovative biotech applications.

Current CRISPR-Cas discovery methods rely on either the identification and annotation of the repetitive CRISPR array sequence (repeat-based annotation), or of subtype-specific *cas* genes (*cas*-based annotation). These traditional annotation methods only comprise two-thirds of the available information of a CRISPR-Cas system, neglecting the information of CRISPR leaders. However, repeat-based CRISPR annotation struggles to annotate unusual CRISPR loci, such as short or degenerate CRISPRs. Whereas Cas-based annotations have no utility for annotating the roughly 25% of bacterial and archaeal CRISPRs that are found far away from *cas* genes. This dissertation highlights my scientific contributions to our understanding of CRISPR leaders, develop bioinformatic tools and how the information of leaders can be incorporated into CRISPR detection tools.

At the beginning of my doctoral research, the importance of the CRISPR leader was not clear. It was known to be involved in transcription but how it regulated the evolution of the CRISPR loci was just starting to be uncovered. Previous work performed in the Doudna group and others showed that several leader sequences of CRISPR arrays of *Pseudomonadota* phylum contain a DNA sequence recognized by a host-specific DNA-binding protein (Integration Host Factor - IHF) that arranges the upstream of CRISPR loci for the spacer integration at the first repeat in the array. However, the distribution and variety of sequence motifs in CRISPR leaders were not thoroughly understood. Analysis of CRISPR leader sequences across a broad spectrum of bacterial and archaeal genomes has shed light on the functional diversity within these pivotal regions. Moreover, incorporating data from CRISPR leaders improves the identification and annotation of CRISPR systems in microbial genomes.

Later work in the course of my Ph.D. focused on a different defense system that involves Argonaute proteins. Like CRISPRs, Argonautes proteins rely on guide nucleic acids to identify complementary foreign nucleic acid. Argonaute (Ago) proteins are found in all three domains of life. In Eukaryotes, Argonautes play a central role in small RNA-guided gene silencing processes collectively known as RNA interference (RNAi). RNAi is involved in many cellular functions such as antiviral defense, gene regulation, DNA repair, alternative splicing, DNA methylation and transposon regulation. Eukaryotic Argonaute (eAgo) proteins make up the catalytic core of the RNAi pathways and are loaded with small interfering RNAs (siRNAs) or microRNAs (miRNAs) that act as guides to identify RNA targets through sequence complementarity. This leads to either target cleavage through intrinsic endonuclease activity, or to silencing of the RNA target by blocking translation. Members of the Argonaute protein family are also present in prokaryotes. A recent comprehensive genomic and phylogenetic analysis identified prokaryotic Argonaute (pAgo) proteins in 17 % of bacterial and 25 % of archaeal genera. The prokaryotic Argonautes are more diverse compared to the eukaryotic Argonautes.

Prokaryotic Argonautes (pAgos) are broken up into three distinct clades based on the domains they possess: long-A, long-B, and short pAgos. Similar to eukaryotic Argonautes (eAgos), long-A pAgos have a bilobed architecture consisting of four domains separated by two linkers: N-terminal, L1 (Linker 1), PAZ (PIWI-Argonaute- Zwiller), L2 (Linker 2), MID (Middle) and PIWI (P-element induced Wimpy testis). Long-B pAgos contain the same domains as Long-A pAgos but have mutations in their PIWI domains that render their nuclease inactive. Short pAgos possess only the MID and PIWI domains and, similar to long-B, have no active nuclease. However, many pAgos with inactive nucleases are linked to genes containing putative nucleases.

Characterized pAgos predominantly fall into the category of long pAgos; however, this overlooks the significant diversity present among short pAgos. Previous studies have identified a proposed “analog of PAZ” (APAZ) domain that is present in many short pAgos as well as in neighboring genes to short pAgos that harbor various effector domains unique to the short pAgo clade. This diversity in the short pAgos highlights a potential sensor/effector mechanism, where the MID and PIWI domain containing short pAgos has a role as sensor that activates the APAZ-effectors. In the last chapter, I have systematically investigated the diversity of pAgos, especially short pAgos, identify flanking APAZ genes and predict the function of various APAZ-Effectors. Large-scale bioinformatic analysis of pAgos and their associated APAZ genes unravels previously undermined APAZ effectors that expands the repertoire of pAgos.

Chapter 1 provides an overview of the significance and mechanisms of the CRISPR-Cas and Argonaute systems. It highlights the revolutionary roles of these systems in genetic regulation and manipulation. The chapter emphasizes the essential function of RNA guides in ensuring precise and targeted genetic interactions in both systems. Further, it provides a short review on RNA-mediated regulation, focusing on the roles of miRNAs and siRNAs associated with Argonautes and the CRISPR-Cas-mediated defense mechanism.

Chapter 2 of this dissertation introduces the type I-F CRISPR-Cas system of *Pseudomonas aeruginosa* and its role in providing immunity against DNA phages. This chapter is pivotal as it examines a hypothesized RNA-targeting mechanism within the type I-F system, which could significantly expand our understanding of CRISPR-Cas versatility beyond DNA interference. Experiments detailed in this chapter aim to test this novel RNA-targeting mechanism by investigating the interactions between the crRNA-guided surveillance complex and RNA targets

of MS2 phage genome. The results indicate a specificity for DNA targets, suggesting that while RNA-targeting by type I-F systems has potential, it may not be a prevalent defense strategy against RNA phages.

Chapter 3 investigates the distribution and phasing of sequence motifs in I-F, I-E, I-C and II-C CRISPR leaders and shows that these motifs are critical to new sequence adaptation. Using bioinformatic and phylogenetic approaches, we show that many CRISPR loci have leaders with sequence motifs that are recognized by Integration Host Factor (IHF) proteins. These leaders additionally contain upstream sequence motifs that had not been previously reported. In the I-E system of *Escherichia coli*, IHF bends the leader DNA into a horseshoe shape during adaptation, facilitating the recruitment of upstream sequence motifs that are recognized by Cas1-2 integrase. Since the spacing of leader motifs are conserved, we hypothesized that the phasing of upstream leader motifs would influence how these motifs are presented to and recognized by CRISPR-associated adaptation proteins. *In vitro* spacer integration assays support these predictions, since perturbations to the conserved spacing between CRISPR leader motifs reduces the ability of Cas1-2 to integrate new spacers. Additionally, we observed that more than half of the type I-E CRISPR leaders did not have IHF-binding sites, but instead had novel DNA sequence motifs. We realized that all the systems with predicted IHF binding sites are found in the bacterial phylum *Pseudomonadota*, which encode for IHF protein. Whereas these CRISPR leaders lack IHF binding sites come from the phylum *Actinomycetota*, which indicates that leader sequences might co-evolve with CRISPR-Cas systems and host factors.

Chapter 4 focusses on the diversity of “leader” sequences of CRISPRs and how the conserved sequence motifs in these regions can be used in enhancing detection capabilities. An

expansive dataset, comprising leaders of 24,940 CRISPRs sourced from 25,900 microbial genomes, serves as the cornerstone for the study. Through rigorous analyses, we investigated conserved sequence motifs in the leaders of CRISPRs, underpinning the intricate dynamics of CRISPR regulation. The chapter methodically categorizes these leader sequences, employing a dual-layered approach: one grounded in 16S rRNA-based taxonomy and the other rooted in CRISPR subtyping. Throughout this systematic classification, we identified 77 distinct motifs, spanning types I, II, III, and V CRISPR leaders. A notable revelation is the alignment of several of these leader motifs with sequences previously implicated in CRISPR-centric processes such as integration, transcription, and processing. This resemblance suggests a regulatory role for the leader motifs within the CRISPR biological framework. The research presented within the chapter also showcases the instrumental role of the identified leader motifs in refining CRISPR detection and classification methods. These motifs emerge as powerful tools for the detection and annotation of CRISPRs.

Chapter 5 critically examines the controversial findings of van Riet et al., who reported the existence of CRISPR-like sequences within the human genome, referred to as "hCRISPRs." Employing a range of bioinformatic tools and methodologies, we aimed to investigate the presence of CRISPRs in the human genome. Initially, we utilized a systematic exploration of the human reference genome using several CRISPR detection tools. Our findings initially suggested the presence of numerous CRISPR-like sequences, partially aligning with those reported by van Riet et al. However, subsequent analyses highlighted inconsistencies when compared to the well-characterized CRISPR systems in bacteria and archaea. To assess the validity of these arrays, we applied confidence scores generated by these tools. A mere fraction of the detected arrays was

deemed of high confidence, a total of 72 putative loci, leading to further scrutiny of the repeat-spacer-repeat CRISPR structure. A subsequent comparison of the lengths of identified repeat sequences against established bacterial and archaeal CRISPRs displayed significant variations, raising further questions about the legitimacy of the asserted hCRISPRs. Despite utilizing various bioinformatic tools, no substantial matches to *cas* genes within the human genome were found. Upon extending our investigation to the comprehensive Telomere-to-Telomere assembly, we identified numerous potential CRISPR arrays. However, none displayed the characteristic features of a true CRISPR system, casting doubt on the existence of "hCRISPRs" in the human genome.

Chapter 6 provides a comprehensive examination of the complexities of Argonaute proteins, with a focus on the less-explored prokaryotic Argonautes. While eukaryotic Argonautes play a well-established role in RNA interference (RNAi), prokaryotic Argonautes remain poorly defined. We applied bioinformatics methods to investigate the sequence diversity short prokaryotic Argonautes (pAgos) and their associated APAZ genes. Preliminary insights reveal that these proteins not only exhibit unique genetic associations, but they also have strong affiliations with various APAZ-effectors. These data suggest short pAgos play a role in immune defense in prokaryotes through diverse mechanisms. Through this in-depth exploration, Chapter 6 aspires to bridge the existing knowledge gap surrounding short pAgos, elucidating their significance in prokaryotic defenses and underscoring their potential for innovative biotechnological applications.

CHAPTER ONE

INTRODUCTION

In the tapestry of molecular biology, certain discoveries emerge as threads of gold, illuminating new paths and reshaping the landscape. Prominent among these are the CRISPR-Cas and Argonaute systems. The CRISPR-Cas system, originally an adaptive bacterial immune mechanism, has transformed into a groundbreaking tool for gene editing. Its profound impact was acknowledged with the award of the Nobel Prize in Chemistry to Emmanuelle Charpentier and Jennifer Doudna for their pioneering work¹. Similarly, the Argonaute proteins, which lie at the heart of RNA interference, emerged as monumental figures in genetic regulation and manipulation, with Andrew Z. Fire and Craig C. Mello being honored with the Nobel Prize for elucidating its mechanisms^{2, 3}. Central to the success and versatility of both the CRISPR-Cas and Argonaute systems is the role of RNA guides, ensuring targeted and specific genetic interactions, underscoring the transformative power of RNA in shaping the future of genetic research and applications.

RNA Guides: Achieving Target Recognition Through CRISPR and Argonautes

RNA plays a pivotal role within the cell as it orchestrates a multitude of mechanistically unique processes⁴. Transfer RNAs (tRNAs) involved in the process of protein synthesis, while small nuclear RNAs (snRNAs) have key roles in the splicing of messenger RNA (mRNA). Additionally, small nucleolar RNAs (snoRNAs) serve as guides for RNA modifications. Furthermore, there are numerous and diverse classes of non-coding RNAs, which play crucial roles

in directing gene regulation or defending the genome to maintain its integrity through a several features^{4, 5}.

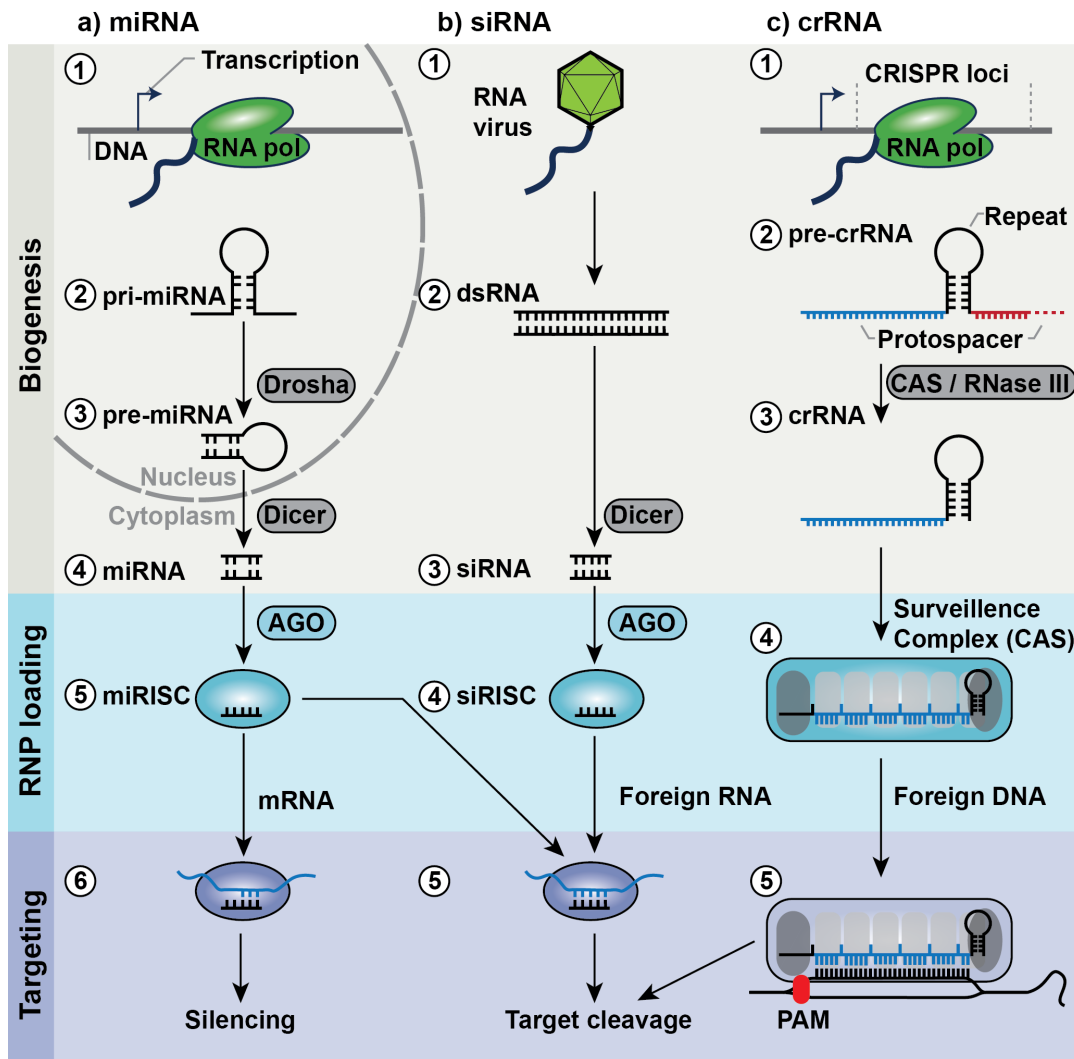


Figure 1.1. The biogenesis, loading and targeting of RNA in distinct RNA-based systems.

(a) MicroRNAs (miRNAs) are generally transcribed by RNA polymerase II (step 1) to produce primary miRNAs (pri-miRNAs) (step 2); these are cleaved by Drosha in the nucleus to produce a stem-loop structure called the pre-miRNA (step 3), which is exported to the cytoplasm and processed by Dicer to form the mature miRNA duplex (step 4), a single strand of which is loaded onto an Argonaute (AGO) protein as part of the miRNA-induced silencing complex (miRISC) (step 5). miRISC binds to target mRNAs through partial complementary base-pairing (step 6), which leads to changes in mRNA stability and/or translational repression or, in cases of high complementarity mRNA cleavage. **(b)** Small interfering RNAs (siRNAs) are generated from

double-stranded RNA (dsRNA) precursors (step 1 and 2), which are cleaved by Dicer into siRNA duplexes (step 3) and loaded onto AGO proteins as part of the siRISC (step 4). The siRISC binds to target mRNAs by highly complementary base-pairing (step 5), which leads to transcript cleavage. **(c)** The CRISPR RNA (crRNA) locus contains repeat sequences that are interspersed with sequences from foreign DNA. It is transcribed by RNA polymerase (step 1) to produce a pre-crRNA. The pre-crRNA is processed by Cas proteins or host ribonucleases to produce the mature crRNA (step 3). The mature crRNA is loaded into surveillance complex (Cas proteins) with nuclease activity (step 4), from which it binds to a complementary foreign DNA containing a protospacer-adjacent motif (PAM; red) (step 5), resulting in DNA cleavage and destruction (not shown).

One of the key functions of RNA involves its ability to regulate nucleic acids in a sequence-specific manner by base-pairing with specific RNA or DNA targets. It can also be rapidly synthesized or processed in response to a signal and it is functional without the need to be translated. Moreover, it has been observed that regulatory RNA molecules have the ability to undergo co-degradation alongside their target molecules upon binding, hence facilitating prompt reactions to a change in cellular status. These responses possess unique quantitative attributes that differentiate them from the protein-based mechanisms of cellular regulation^{5, 6}. In the vast landscape of RNA-mediated regulation, two systems stand out as pioneers in RNA-based target recognition: the miRNA and siRNAs, intricately associated with Argonautes (Ago), and the formidable CRISPR-Cas-mediated defense mechanism (**Figure 1.1**).

RNA-based Recognition by miRNAs and siRNAs in Argonautes

The discovery of micro RNAs (miRNAs) and small interfering RNAs (siRNAs) revolutionized the idea that RNA is a regulatory molecule in eukaryotic organisms. Computational and experimental approaches suggest that eukaryotic genomes encode hundreds of miRNAs^{7, 8}, and it is estimated that the most conserved miRNAs control the expression of approximately two-thirds of human genes⁹. Although the origin and precursors of miRNAs and siRNAs differ, they

share in part common biogenesis pathways and they are both loaded onto Argonaute (Ago) proteins, which they guide to mRNA targets for RNA silencing^{10, 11}.

miRNAs are sRNAs of ~21–23 nucleotides in length that are encoded by the genome and some viruses. The majority of miRNAs originate from independent loci, although some are processed from within introns. Their primary transcripts contain stem–loop structures, and consecutive endonuclease cleavages by two enzymes belong to RNase III protein family, Drosha in the nucleus and Dicer in the cytoplasm, together generate a miRNA duplex containing a 5' monophosphate and a 3' hydroxyl. The mature miRNA strand is loaded into the Ago protein within the RNA-induced silencing complex (RISC)^{11, 12} (**Figure 1.1a**). The miRNA within RISC associates with target mRNAs through base-pairing, leading to translational repression and/or mRNA destabilization¹³ or, in some cases when the miRNA is highly complementary to its target, mRNA cleavage (**Figure 1.1a, b**).

siRNAs differ from miRNAs in that they are excised from long, fully complementary, double-stranded RNAs; these can originate from transgenes, viral infections and repetitive elements¹¹. After the siRNA precursors are cleaved by Dicer into duplexes of ~21 nucleotides they are loaded into the RISC (**Figure 1.1b**), from which they recognize and associate with target transcripts through base-pairing. In both miRNA and siRNA carrying RISC, the base pairing starts with the recognition of seed region consists of eight nucleobases. The seed region facilitates the base pairing by pre-ordering these nucleobases in a helical configuration which gives this part of the sequence a thermodynamic advantage for interacting with the target sequence. This complementary between the seed region and target enables to investigate the rest of the target,

which leads to the endonucleolytic cleavage of the complementary sequence by Ago within RISC as a result of high levels of complementarity¹².

Initial insights into the molecular basis behind the importance of the seed sequence were obtained from crystal structures of the bacterial Ago protein from *Thermus thermophilus* (TtAgo). Even though these structures used guide DNAs, they revealed how guide strands mediate RNA recognition¹⁴⁻¹⁶ (**Figure 1.2a**). TtAgo containing a 5'-phosphorylated 21-nucleotide DNA guide strand maintained the same overall bi-lobal structure containing four domains: MID, PIWI, N and PAZ domains (**Figure 1.2a**).

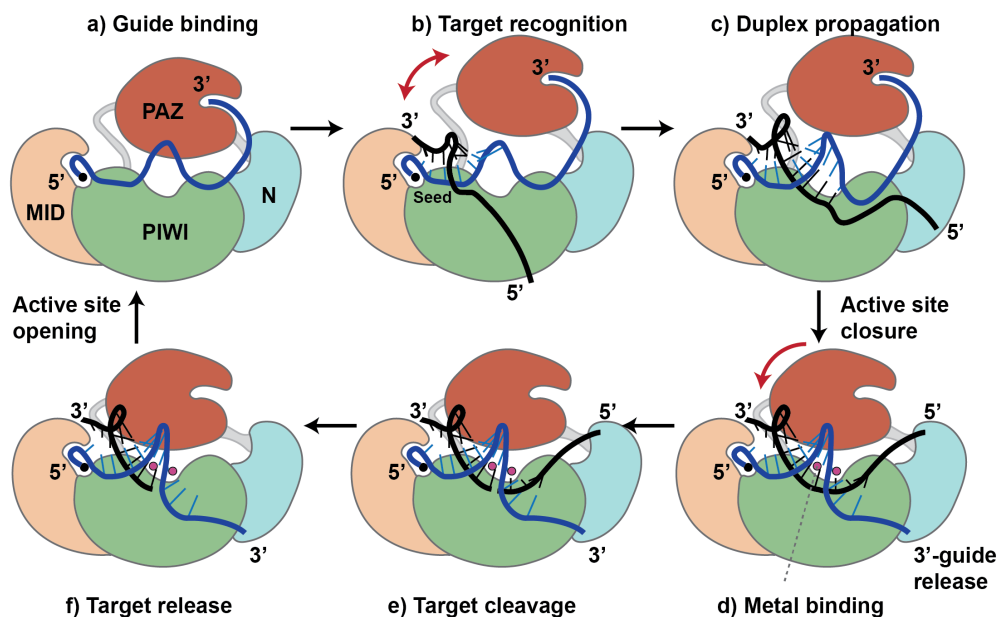


Figure 1.2. Presentation of seed sequences in guide strands and target cleavage by Ago. **(a-b)** Guide-loaded Ago performs search for a complementary target through base-pairing with the seed region of the guide strand, followed by duplex propagation through the central part and the 3'-supplementary site of the guide. **(c-d)** Conformational mobility of the PAZ domain (shown by arc-shaped arrows) facilitates correct base-pairing, through controlled release of the guide 3'-end and active site closure. **(e-f)** Conformational changes in the active site allow binding of catalytic metal ions, followed by cleavage of the target strand and its stepwise release from the complex. The guide strand is blue, the target strand is black; only the target strand of DNA substrate is shown.

The DNA guide is nestled within a basic channel spanning the lobes that harbor the PAZ domain and the PIWI domain, and there are extensive contacts between its backbone phosphates and basic residues in the domains and linkers present in the protein.¹⁷ The trajectory of the guide is defined by the anchoring of its 5' monophosphate into a pocket within the MID domain and of its 3' end in the PAZ domain. The conformation of the seed sequence at nucleotides 2–6 reveals how its presentation within the TtAgo ribonucleoprotein (RNP) is important for RNA recognition. Specifically, in agreement with earlier biochemical and biophysical studies¹⁸, these nucleotides are stacked forming an A-form helix with their Watson-Crick edges exposed to the solvent, a favorable, pre-organized conformation for interacting with a target mRNA (**Figure 1.2a-c**). Subsequent nucleotides are threaded through a narrow channel in the TtAgo protein and are therefore inaccessible to mRNA. Introducing a kink into the guide strand by positioning two Arginine residues between nucleotides 10 and 11 renders the binary TtAgo protein–guide complex catalytically inactive and prevents the formation of the undistorted helical conformation that is required for mRNA cleavage^{14, 15, 17}. However, on the addition of a 20-nucleotide target RNA to the TtAgo protein–guide complex, the seed region forms an A-form DNA–RNA helix with the target and undergoes a conformational change that relieves the 3' end of the guide from PAZ domain, enabling cleavage of the target mRNA^{14, 15, 17} (**Figure 1.2d-f**). The importance of mRNA–seed sequence complementarity is highlighted by the fact that cleavage is impaired on the introduction of mismatches or insertions within the seed region in the guide DNA¹⁵. Therefore, the pre-organization of the seed in an A-form helix within TtAgo facilitates its interaction with the target mRNA and offsets the entropy penalty of binding target mRNA, enhancing the interaction affinity up to 300-fold¹⁹.

RNA-based Recognition by crRNAs in CRISPR-Cas Systems

Many bacteria and most archaea contain an RNA-mediated interfering mechanism that provides adaptive immunity against bacteriophages and plasmids²⁰. This mechanism uses CRISPR RNAs (crRNAs) generated from genomic CRISPR loci for the sequence-specific detection and elimination of invading viruses and plasmids. crRNA sequences originate from previous invading nucleic acids and act as guides for various CRISPR-associated Cas proteins, thus providing an inherited adaptive immunity against re-infections. CRISPR loci can be classified into six main systems, depending on their Cas protein composition²¹, but all contain a nuclease and a guide RNA. Type I, type III and type IV systems (Class I) use crRNA that is incorporated into a multi-subunit complex to execute their interference activity while type II, type V and type VI systems (Class II) contains single Cas protein to function as RNA-guided endonucleases. This chapter focus on the type I-F system to explain the mechanism of RNA recognition by the Csy (surveillance) complex and cleavage achieved by Cas2/3 endonuclease.

CRISPR–Cas adaptive immunity consists of three phases: immunization, where new spacer sequences from invading nucleic acids are incorporated into the first repeat of CRISPR array (depicted by CRISPR leader) to function as a molecular memory of prior infections; expression, which involves the transcription and processing of crRNAs; and interference, whereby the crRNAs are used as guides to identify foreign nucleic acids and facilitate their cleavage and clearance by associated nucleases²⁰.

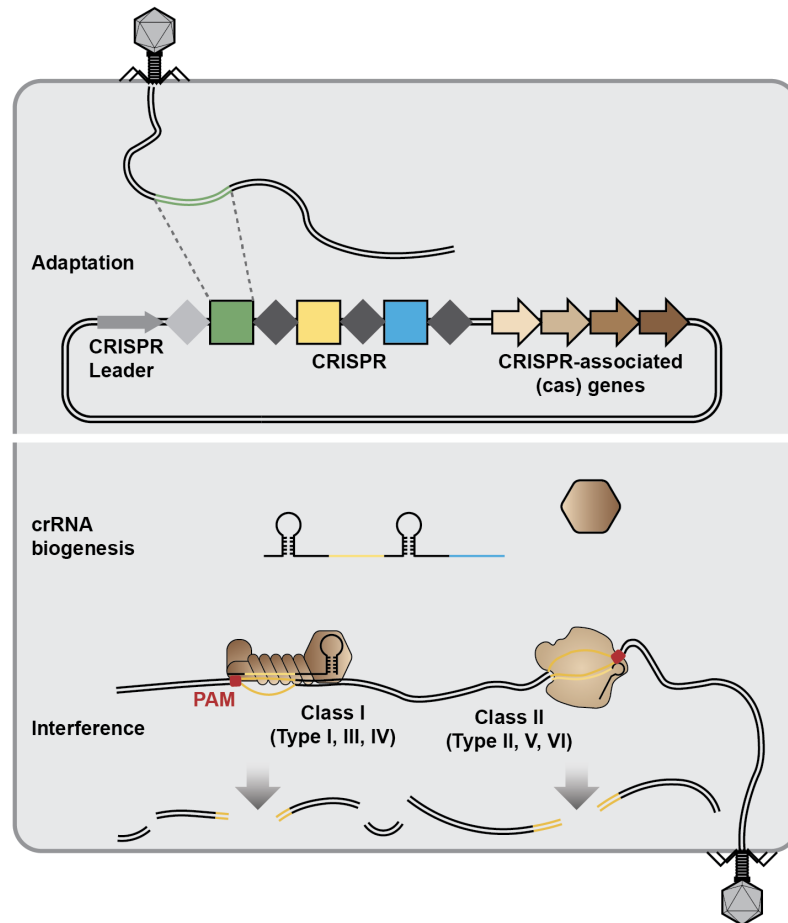


Figure 1.3. The key steps of CRISPR-Cas immunity.

Adaptation: insertion of new spacers into the CRISPR locus. crRNA biogenesis: transcription of the CRISPR locus and processing of CRISPR RNA. Interference: detection and degradation of mobile genetic elements by CRISPR RNA and Cas protein(s).

RNA plays a fundamental part in the expression and interference phases, where the CRISPR loci is transcribed into pre-crRNA containing repeats and spacers. The subsequent processing of this transcript achieved by Cas6 protein²² (**Figure 1.1c and 1.3**). The resulting mature crRNA contains 8 nucleotide partial repeat, 32 nucleotide spacer-derived guide sequence and a 20-nucleotide repeat sequence with the stem-loop structure²³. Six copies of Cas7 proteins form a backbone that covers crRNA with Cas5 and Cas8 attached to the 5' of crRNA. The unique

hand-like structure of the Cas7 protein interacts with the crRNA, introducing two consecutive $\sim 90^\circ$ turns in its backbone^{24, 25}. These chicanes in the crRNA occur at a regular 6-nt periodicity; each chicane is separated by five bases presented to the solvent in pseudo A-form, while the sixth base is flipped out and covered by the thumb of an adjacent Cas7 molecule^{23, 25}. This pre-organized helical conformation places the crRNA in a configuration that is thermodynamically favorable for guide–target duplex formation.

Identifying targets is a stepwise process involving not only base-pairing between DNA and the crRNA sequence but also recognition of a short sequence motif called the protospacer-adjacent motif (PAM) on the non-complementary strand of the DNA by the surveillance complex²⁶. The PAM is required for RNP to recognize and stably bind to DNA (**Figure 1.4a, b**); for example, a 5'-NGG sequence (where 'N' is any nucleobase) immediately flanking the protospacer. PAM recognition induces local distortion of the DNA duplex that facilitates crRNA-guided hybridization to the complementary DNA target, which displaces the non-complementary DNA strand (R-loop)²⁶. This recognition of the PAM sequence and the A-form helical arrangement of the seed sequence (8 nucleotide followed after PAM sequence) are coupled to conformational changes that enable surveillance complex to recognize targets while results in an elongation of the Cas7 backbone (**Figure 1.4c**)²⁷. The elongated conformation also creates a gap between the head and the tail of the complex that is necessary for a $\sim 180^\circ$ rotation of the helical bundle of Cas8 (**Figure 1.4d**). This rotation leads to recruitment of the trans-acting nuclease/helicase Cas2/3 for the target DNA degradation^{23, 28} (**Figure 1.4e, f**). Similar to Agos, the complementarity between target DNA and seed sequence is crucial, where mismatches or between the seed region and the target DNA impair cleavage²⁷.

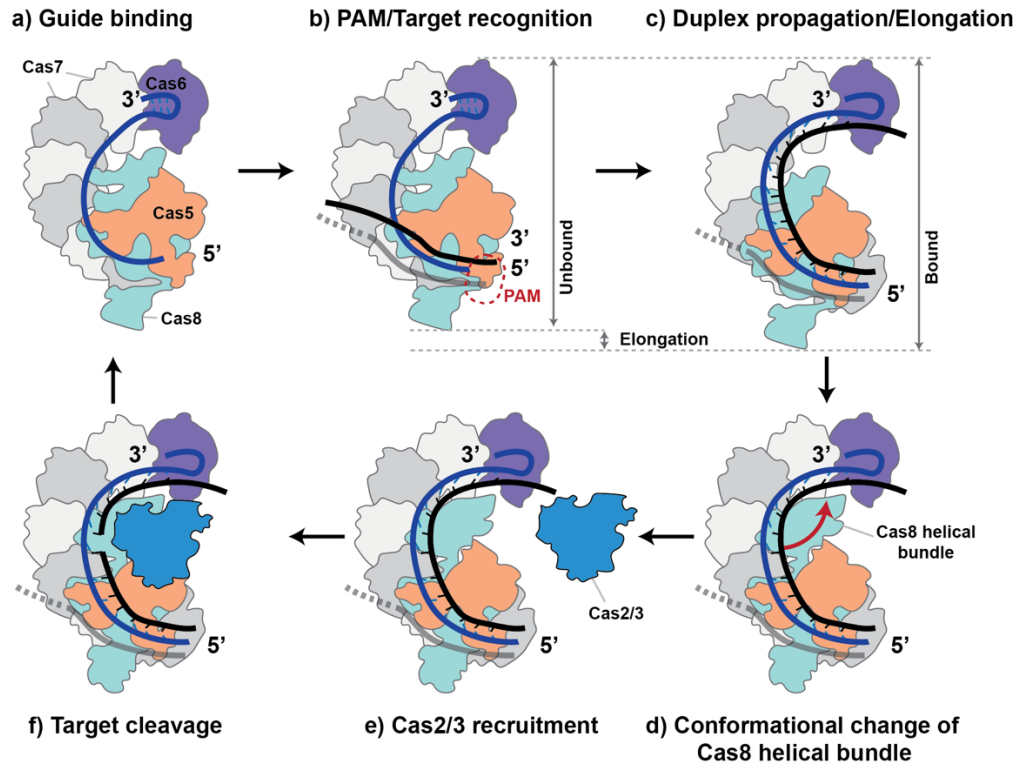


Figure 1.4. Presentation of target binding in guide strands, conformational change of surveillance complex (Csy complex) and target cleavage by recruited Cas2/3 trans-acting nuclease/helicase. **(a-b)** Guide-loaded Csy complex performs search for a 5'-NGG-3' PAM and complementary target through base-pairing with the seed region of the guide strand, followed by duplex propagation through the central part and the 3'-supplementary site of the guide. **(c-d)** Conformational mobility of the Cas7 backbones creates a gap between the head and the tail of the complex that is necessary for a $\sim 180^\circ$ rotation of the helical bundle of Cas8. **(e-f)** Conformational changes of the Cas8 helical bundles recruits Cas2/3 trans-acting nuclease/helicase followed by cleavage of the target strand. The guide strand is blue, the target strand is black.

Summary

A combination of structural, biochemical, biophysical, and genome-wide approaches has provided extensive insights into how short RNA sequences can identify targets in the cell^{1, 3, 17, 20}. This has revealed shared principles, which are especially clear for the miRNA, siRNA, and crRNA utilizing regulatory/defense systems. Moreover, a need has risen to investigate the diversity of these systems with a possibility to identify different candidates that can be used as biotechnology

tools²¹. In this dissertation, I embarked on a comprehensive analysis of the CRISPR-Cas systems, emphasizing their molecular mechanisms, evolutionary patterns, and interactions within microbial genomes. Through a synthesis of experimental results and bioinformatic analyses, I shed light on the crucial role of sequence motifs in CRISPR leader sequences and their impact on adaptation and immunity. Furthermore, I critically examined the controversial presence of CRISPR loci in the human genome, providing clarity on its legitimacy. Concluding the journey, I systematically investigated into the lesser-known world of prokaryotic Argonaute proteins, unveiling their potential significance in microbial defense systems and potential biotechnological applications.

References

1. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity*. Science, 2012. **337**(6096): p. 816-21.
2. Fire, A., et al., *Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans**. Nature, 1998. **391**(6669): p. 806-11.
3. Zamore, P.D., *RNA interference: big applause for silencing in Stockholm*. Cell, 2006. **127**(6): p. 1083-6.
4. Cech, T.R. and J.A. Steitz, *The noncoding RNA revolution-trashing old rules to forge new ones*. Cell, 2014. **157**(1): p. 77-94.
5. Levine, E. and T. Hwa, *Small RNAs establish gene expression thresholds*. Curr Opin Microbiol, 2008. **11**(6): p. 574-9.
6. Masse, E., F.E. Escorcia, and S. Gottesman, *Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli**. Genes Dev, 2003. **17**(19): p. 2374-83.
7. Chiang, H.R., et al., *Mammalian microRNAs: experimental evaluation of novel and previously annotated genes*. Genes Dev, 2010. **24**(10): p. 992-1009.
8. Fromm, B., et al., *A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome*. Annu Rev Genet, 2015. **49**: p. 213-42.
9. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs*. Genome Res, 2009. **19**(1): p. 92-105.
10. Meister, G., *Argonaute proteins: functional insights and emerging roles*. Nat Rev Genet, 2013. **14**(7): p. 447-59.
11. Czech, B. and G.J. Hannon, *Small RNA sorting: matchmaking for Argonautes*. Nat Rev Genet, 2011. **12**(1): p. 19-31.
12. Kim, V.N., J. Han, and M.C. Siomi, *Biogenesis of small RNAs in animals*. Nat Rev Mol Cell Biol, 2009. **10**(2): p. 126-39.
13. Jonas, S. and E. Izaurralde, *Towards a molecular understanding of microRNA-mediated gene silencing*. Nat Rev Genet, 2015. **16**(7): p. 421-33.
14. Wang, Y., et al., *Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes*. Nature, 2009. **461**(7265): p. 754-61.
15. Wang, Y., et al., *Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex*. Nature, 2008. **456**(7224): p. 921-6.

16. Wang, Y., et al., *Structure of the guide-strand-containing argonaute silencing complex*. Nature, 2008. **456**(7219): p. 209-13.
17. Lisitskaya, L., A.A. Aravin, and A. Kulbachinskiy, *DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins*. Nat Commun, 2018. **9**(1): p. 5165.
18. Haley, B. and P.D. Zamore, *Kinetic analysis of the RNAi enzyme complex*. Nat Struct Mol Biol, 2004. **11**(7): p. 599-606.
19. Parker, J.S., et al., *Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein*. Mol Cell, 2009. **33**(2): p. 204-14.
20. Marraffini, L.A., *CRISPR-Cas immunity in prokaryotes*. Nature, 2015. **526**(7571): p. 55-61.
21. Makarova, K.S., et al., *Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants*. Nat Rev Microbiol, 2020. **18**(2): p. 67-83.
22. Wakefield, N., R. Rajan, and E.J. Sontheimer, *Primary processing of CRISPR RNA by the endonuclease Cas6 in Staphylococcus epidermidis*. FEBS Lett, 2015. **589**(20 Pt B): p. 3197-204.
23. Rollins, M.F., et al., *Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry*. Mol Cell, 2019. **74**(1): p. 132-142 e5.
24. Hrle, A., et al., *Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3*. RNA Biol, 2013. **10**(11): p. 1670-8.
25. Lintner, N.G., et al., *Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE)*. J Biol Chem, 2011. **286**(24): p. 21643-56.
26. Rollins, M.F., et al., *Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa*. Nucleic Acids Res, 2015. **43**(4): p. 2216-22.
27. Wiedenheft, B., et al., *RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions*. Proc Natl Acad Sci U S A, 2011. **108**(25): p. 10092-7.
28. Rollins, M.F., et al., *Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity*. Proc Natl Acad Sci U S A, 2017. **114**(26): p. E5113-E5121.

CHAPTER TWO

TYPE I-F CRISPR-CAS PROVIDES PROTECTION
FROM DNA, BUT NOT RNA PHAGES

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Murat Buyukyoruk

Contributions: Planned and performed the experiments, analyzed the data, and wrote the manuscript.

Co-Author: Blake Wiedenheft

Contributions: Planned and oversaw the experiments, analyzed the data, and wrote the manuscript.

Manuscript Information

Murat Buyukyoruk and Blake Wiedenheft

Cell Discovery

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- X Published in a peer-reviewed journal

Published by Springer Nature

Volume 5

Number 54

22 October 2019

DOI: 10.1038/s41421-019-0123-9

TYPE I-F CRISPR-CAS PROVIDES PROTECTION FROM DNA, BUT NOT RNA
PHAGES

Murat Buyukyoruk¹, Blake Wiedenheft¹

¹Department of Microbiology and Immunology, Montana State University, Bozeman MT
59717, USA.

Manuscript

The type I-F CRISPR-Cas system from *Pseudomonas aeruginosa* (PA14) provides sequence specific elimination of invading DNA^{1, 2}. Recently, a paper published in *Cell Research* reported that this system targets mRNA through a non-canonical mechanism³. Here, we implement the proposed design rules for generating CRISPR-RNA (crRNA)-guides that target RNA and test whether the proposed RNA-targeting activity provides immunity against RNA phage infection. Our experiments reveal that the type I-F CRISPR system provides protection from DNA, but not RNA phages.

The proposed mechanism of RNA targeting is distinct from dsDNA targeting. Rather than detection of a dsDNA PAM via protein mediated recognition of specific nucleobases^{4, 5}, Li *et al.* suggest that single-stranded RNA recognition proceeds via recognition of a “PAM-like” motif (5'-GGN-3') on the opposite side of the complex and that this is followed by detection of a 12-nt “core” sequence (**Figure 2.1a**)³. To assess the versatility of these RNA targeting rules, we designed I-F CRISPRs to target the RNA genome of MS2 phage⁶.

We designed three synthetic CRISPR arrays that each contain spacers designed to target six different regions in the MS2 RNA phage genome (**Figure S2.1**). Spacers in the first CRISPR (CRISPR-1) were designed to target sequences with a “PAM-like” NGG motif, a “core” sequence, and mismatches along the target sequence identical to those identified by Li *et al.*³ (**Figure 2.1b**). Spacers in the second CRISPR (CRISPR-2), were designed to target the same RNA sequences, but the crRNA-guide is 100% complementary to the RNA target (i.e., no mismatches). Spacers in the third CRISPR (CRISPR-3) were designed according to established dsDNA targeting rules. Secondary structures in the MS2 genome are likely to be important for RNA targeting, so the six

spacers included in each CRISPR were designed to target locations with varying degrees of predicted secondary structure. These structures are similar to secondary structures predicted for the *LasR* mRNA, which is purported to be targeted for degradation by the type I-F CRISPR system in PA14³.

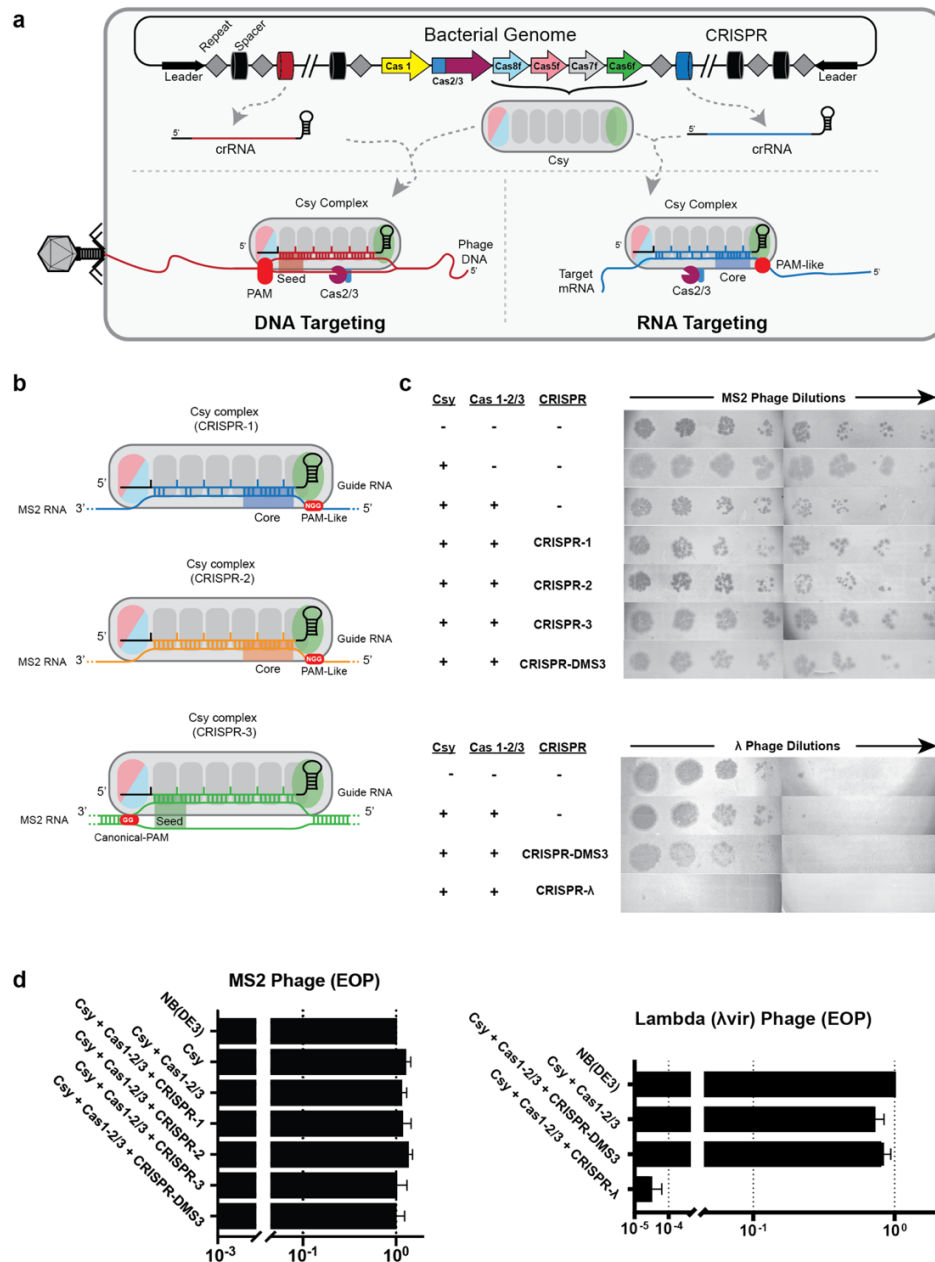


Figure 2.1. Testing the efficiency of type I-F CRISPR defense against RNA and DNA phages.

(a) The type I-F CRISPR-Cas system of *P. aeruginosa* (PA14) contains six *cas* genes flanked by two CRISPR loci. DNA binding by the crRNA-guided surveillance complex (Csy complex) relies on recognition of a double-stranded PAM followed by a complementary protospacer. Base pairing occurs in discrete increments of five. Every sixth nucleobase is covered by a beta-hairpin on Cas7 and these bases do not contribute to target interactions^{4, 5}. Target binding drives a conformational change in the Csy complex that recruits the Cas2/3 nuclease for target cleavage. In contrast to dsDNA binding, the proposed RNA-binding model requires the recognition of a single-stranded “PAM-like” sequence followed by a 12-nt core sequence. **(b)** Schematics of CRISPRs 1, 2, and 3. CRISPR-1 (top) represents the model proposed by Li *et al.*³. CRISPR-2 (middle) targets the same sequences as CRISPR-1, but the crRNA-guide is completely complementary. CRISPR-3 (bottom) was designed using well-established rules for dsDNA-binding mechanism. **(c)** Ten-fold of dilutions of MS2 phage (top) or Lambda (λ) phage (bottom) on a lawn of wild-type *E. coli* NB(DE3) cells or NB(DE3) expressing the type I-F CRISPR system. **(d)** The efficiency of plaquing (EOP) of MS2 phage (left) and lambda phage (right). Efficiency of plaquing (EOP) was calculated as the ratio of plaque forming unites (PFUs) on I-F CRISPR expressing cells divided by the number of PFUs on control cells.

To measure the efficiency of phage defense, we performed a series of plaque assays (**Figure 2.1c**). None of the CRISPRs designed to target MS2 provided protection. To confirm that our experimental system is functional, we replaced the MS2 targeting CRISPR, with a CRISPR designed to target Lambda (λ) phage (CRISPR- λ). Cells that express Csy, Cas2/3 and a synthetic CRISPR designed to target λ -phage, resulted in a 5-log reduction of plaques, as compared to non-targeting controls (i.e., CRISPR-DMS3) (**Figure 2.1d**).

Importantly, the work presented here was not designed to replicate the previously published work by Li *et al.*, but rather our intention was to determine if the non-canonical crRNA-guided recognition rules describe by these authors could be generally applied for programmable defense against RNA phages. Unlike the type III CRISPR systems described by Silas *et al.*, the type I-F systems do not contain a reverse transcriptase, which would be necessary to incorporate new spacers from RNA-based parasites⁷. Here, we by-pass new spacer acquisition, by creating synthetic CRISPRs designed to target the MS2 genome. Regardless of the design rule (canonical

or non-conical), none of the MS2 targeting CRISPRs were capable of knocking down RNA phage infection. This biological result is supported by recent structural, biochemical, and biological assays which all show that Cas2/3 nuclease recruitment requires a conformational change in the Csy complex that is dependent on displacement of a non-complementary strand during crRNA-guided base pairing to the target DNA⁴. We acknowledge that CRISPR mechanisms are remarkably diverse and we remain open to alternative mechanisms for target recognition and Cas2/3 recruitment, however Høyland-Kroghsbo *et al.* recently repeated the work by Li *et al.* but did not find evidence for RNA targeting in PA14⁸. Collectively, these results suggest that type I-F CRISPR systems are capable of crRNA-guided detection and destruction of dsDNA, but not RNA.

Methods and Materials

Strains, Plasmids and Growth Conditions

NovaBlue (DE3) [*endA1 hsdR17 (r_{K12}- m_{K12}+)* *supE44 thi-1 recA1 gyrA96 relA1 lac (DE3)* *F' proA⁺B⁺ lacIq ZΔM15::Tn10(Tet^R)*] cells were cultured at 37°C in liquid LB or on LB agar, supplemented with 0.1% glucose, 2mM CaCl₂ and 0.01 mg/ml thiamine. Plasmids used for Csy and Cas1-2/3 expression are available at AddGene (89232 and 89240, respectively)^{1,9}. CRISPR-DMS3 is available at AddGene (89244)⁹, while the other CRISPRs were designed, synthesized, and cloned into a PMK vector (**Figure S2.1**). CRISPR-λ contains four spacer sequences that target four Lambda phage genes (J, O, R and E). Growth media is supplemented with ampicillin to maintain the pCys1-4, spectinomycin to maintain pCas1-2/3, and kanamycin to maintain pCRISPR. Additionally, tetracycline is provided in the growth media for the selection of F' episome in NovaBlue (DE3). In NovaBlue (DE3) cells, the induction of expression vectors was achieved on LB Agar plates with 0.001mM isopropyl-D-1-thiogalactopyranoside (IPTG).

Synthetic CRISPR Array Construction

Three different CRISPR arrays were assembled that each contains six spacer sequences to target MS2 phage. Overlapping oligonucleotides (Eurofins Genomics) were annealed and extended at 72°C with Q5 polymerase to synthesize small fragments containing repeat-spacer-repeat. Small fragments were combined using overlap PCR extension and the final CRISPR was cloned using SacI and EcoRI sites (**Figure S2.1**). Each clone was sequenced (Genscript). Each CRISPR contains six spacers (32-bps) flanked by 28-bp repeats. Spacers were selected from a variety of locations on the MS2 genes based on the availability of the PAM or PAM-like region. CRISPRs were designed using the canonical model or non-canonical binding models proposed by Li *et al*³.

Phage Replication and Plaque Assays

E. coli BL21 (DE3) and K12 ER2738 used to propagate Lambda_{vir} and MS2 phages, respectively. K12 ER2738 contains an F' episome under tetracycline selection, which provides receptor for the MS2 phage attachment. 100mls of *E. coli* BL21 (DE3) and K12 ER2738 cells were grown to OD₆₀₀ of 0.5 and then infected with 100µl of corresponding phage. Cultures containing phage were agitated overnight at 37°C. Cell lysates were collected by centrifugation (4000g for 30 minutes). A 0.2µm filter was used to remove particulates from cell lysate. 100mls of the cell lysates were concentrated into 5ml and tittered by plaque assay. Tittering was performed using 100µl of *E. coli* cells grown to OD₆₀₀ of 0.5. These cells were mixed with 100µl of phage or 10-fold dilutions of each phage, incubated for 30 mins and shaken every 10 mins. 5mls of molten 0.8% LB/agarose was added to the cell slurry and the soft agar was poured over the top of an LB-agar plate. Plates were incubated at 37°C overnight and plaques were counted the following day.

Spot Plaque Assay

E. coli NovaBlue (DE3) cells were cultivated with appropriate antibiotics to maintain plasmids and challenge with Lambda_vir (dsDNA) (University of Laval, reference phage HER 37) or MS2 (+ssRNA) phages (ATCC 15597). NovaBlue (DE3) cells were grown to an OD₆₀₀ of 0.9 and 400µl of the bacterial culture is mixed with molten 0.8% LB/agarose and poured over a solid agar. 5µl of each phage dilution were spotted on the top agar and the plates were incubated at 37°C overnight. Plates were evaluated the following day.

Quantification and Statistical Analysis

Efficiency of plaquing (EOP) is reported as a ratio of PFUs formed on the experimental (i.e. CRISPR containing) strain divided by the number of PFUs on a wild-type. Plaque assays were performed in triplicate and the EOP is the $\pm$ standard deviation.

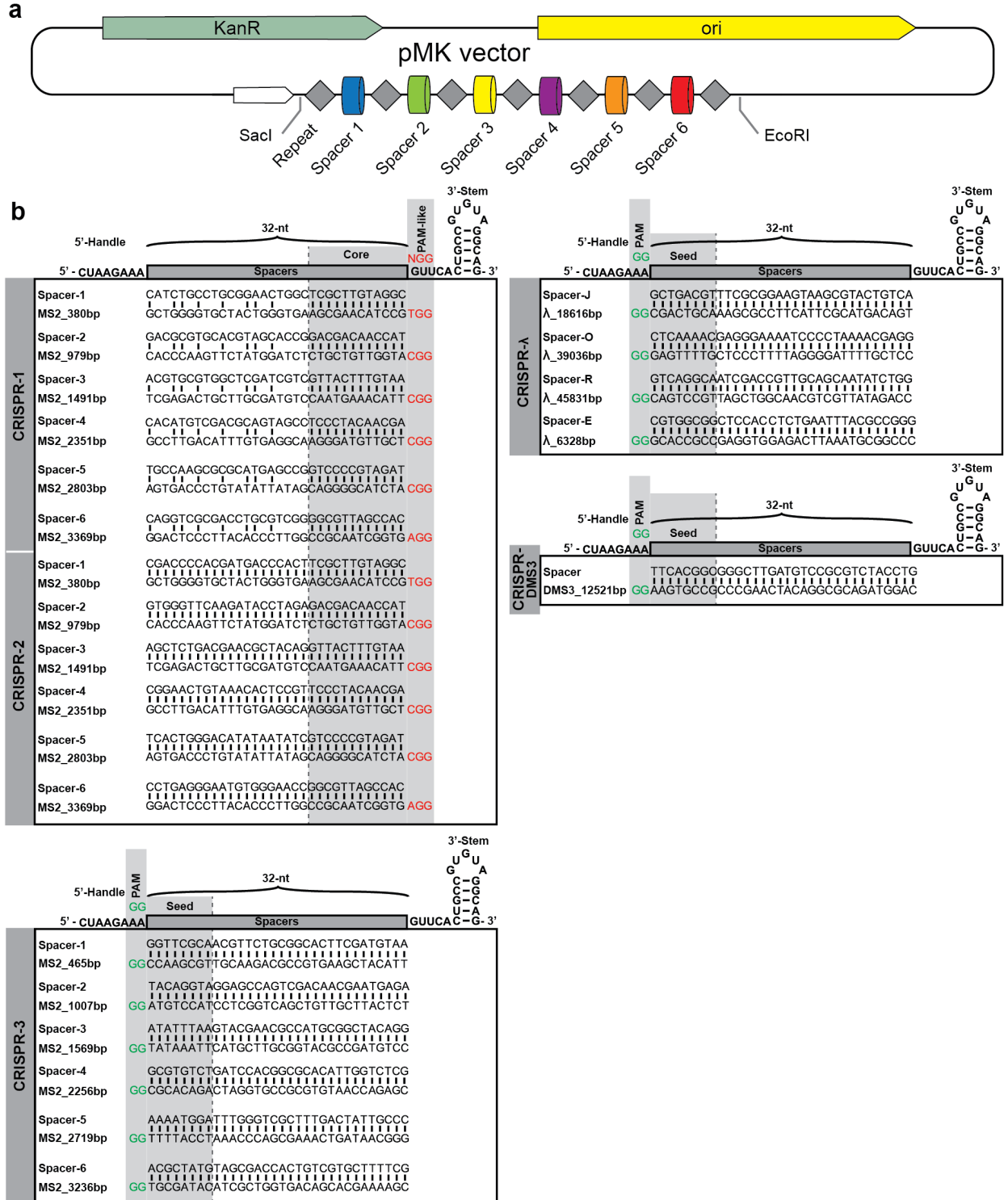
Acknowledgements

We thank members of the Wiedenheft laboratory for comments, suggestions, and critical feedback. Research in the Wiedenheft laboratory is supported by National Institutes of Health (P20GM103500, P30GM110732-03, R01GM110270, and R01GM108888), National Science Foundation (EPSCoR EPS-110134), the M. J. Murdock Charitable Trust, and the Montana Agricultural Experimental Station.

References

1. Rollins, M.F., et al., *CasI and the Csy complex are opposing regulators of Cas2/3 nuclease activity*. Proc Natl Acad Sci U S A, 2017. **114**(26): p. E5113-E5121.
2. Gong, B., et al., *Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3*. Proc Natl Acad Sci U S A, 2014. **111**(46): p. 16359-64.
3. Li, R., et al., *Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity*. Cell Res, 2016. **26**(12): p. 1273-1287.
4. Rollins, M.F., et al., *Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry*. Mol Cell, 2019. **74**(1): p. 132-142 e5.
5. Guo, T.W., et al., *Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex*. Cell, 2017. **171**(2): p. 414-426 e12.
6. Dai, X., et al., *In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus*. Nature, 2017. **541**(7635): p. 112-116.
7. Silas, S., et al., *Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-CasI fusion protein*. Science, 2016. **351**(6276): p. aad4234.
8. Hoyland-Kroghsbo, N.M., K.A. Munoz, and B.L. Bassler, *Temperature, by Controlling Growth Rate, Regulates CRISPR-Cas Activity in Pseudomonas aeruginosa*. mBio, 2018. **9**(6).
9. Chowdhury, S., et al., *Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex*. Cell, 2017. **169**(1): p. 47-57 e11.

Supplemental Figures



cCRISPR-1 array

TAATACGACTCACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGAGCTC**GTTCACTGCCGTGTAGGCA**
GCTAAGAAACATCTGCCTGCGGAAGTGGCTCGCTTGTAGGC**GTTCACTGCCGTGTAGGCAGCTAAGAAA**GACGCGT
GCACGTAGCACCGGACGACAACCAT**GTTCACTGCCGTGTAGGCAGCTAAGAAA**ACGTGCGTGGCTCGATCGTCGTT
ACTTTGTAA**GTTCACTGCCGTGTAGGCAGCTAAGAAA**CACATGTCGACGCAGTAGCCTCCCTACAACGA**GTTCACTG**
CCGTGTAGGCAGCTAAGAAATGCCAAGCGCGCATGAGCCGTTCCCGTAGAT**GTTCACTGCCGTGTAGGCAGCTAA**
GAAACAGGTCGCGACCTGCGTCGGGGCGTTAGCCAC**GTTCACTGCCGTGTAGGCAGCTAAGAAA**

CRISPR-2 array

TAATACGACTCACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGAGCTC**GTTCACTGCCGTGTAGGCAG**
CTAAGAAACGACCCACGATGACCCACTTCGCTTGTAGGC**GTTCACTGCCGTGTAGGCAGCTAAGAAA**GTGGGTTT
AAGATACCTAGAGACGACAACCAT**GTTCACTGCCGTGTAGGCAGCTAAGAAA**GCTCTGACGAACGCTACAGGTTAC
TTTGTAA**GTTCACTGCCGTGTAGGCAGCTAAGAAA**CGGAAGTGAACACTCCGTTCCCTACAACGA**GTTCACTGCC**
GTGTAGGCAGCTAAGAAATCACTGGGACATATAATATCGTCCCGTAGAT**GTTCACTGCCGTGTAGGCAGCTAAGAAA**
CCTGAGGGAATGTGGGAACCGGCGTTAGCCAC**GTTCACTGCCGTGTAGGCAGCTAAGAAA**

CRISPR-3 array

TAATACGACTCACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGAGCTC**GTTCACTGCCGTGTAGGCA**
GCTAAGAAAGGTTGCAACGTTCTGCGCACTTCGATGTAAG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**TACAGGT
AGGAGCCAGTCGACAACGAATGAGA**GTTCACTGCCGTGTAGGCAGCTAAGAAA**ATATTTAAGTACGAACGCCATGC
GGCTACAGG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**GCCTGTCTGATCCACGGCGCACATTGGTCTC**GGTTCAC**
TGCCGTGTAGGCAGCTAAGAAAAAAAATGGATTGGGTGCTTTGACTATTGCC**GTTCACTGCCGTGTAGGCAGCT**
AAGAAAACGCTATGTAGCGACCACTGTCGTGCTTTTCG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**

CRISPR-λ array

GTTCACTGCCGTGTAGGCAGCTAAGAAAGCTGACGTTTCGCGGAAGTAAGCGTACTGTCA**GTTCACTGCCGTGTAG**
GCAGCTAAGAAACTCAAAACGAGGGAAAATCCCCTAAAACGAGG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**GTCA
GGCAATCGACCGTTGCAGCAATATCTGG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**CGTGGCGGCTCCACCTCTG
AATTTACGCCGGG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**CAGGTAGACGCGGACATCAAGCCCGCGTGAAAG
GTTCACTGCCGTGTAGGCAGCTAAGAAA

CRISPR-DMS3 array

TAATACGACTCACTATAGG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**TTACGGCGGGGCTTGATGTCCGCGTCTACC
TG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**TTACGGCGGGGCTTGATGTCCGCGTCTACCT**GGTTCAGTCCGTGT**
AGGCAGCTAAGAAATTACGGCGGGGCTTGATGTCCGCGTCTACCTG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**TT
CACGGCGGGCTTGATGTCCGCGTCTACCTG**GTTCACTGCCGTGTAGGCAGCTAAGAAA**TTACGGCGGGGCTTGATG
TCCGCGTCTACCT**GGTTCAGTCCGTGTAGGCAGCTAAGAAA**TTACGGCGGGGCTTGATGTCCGCGTCTACCT**GGT**
TCACTGCCGTGTAGGCAGCTAAGAAATTACGGCGGGGCTTGATGTCCGCGTCTACCTG**GTTCACTGCCGTGTAGGC**
AGCTAAGAAA

Figure S2.1. CRISPR design.

(a) Plasmid map of pMK vector containing six spacer sequences designed to target the RNA genome of MS2 phage. Each CRISPR array contains six (32-bps) spacers flanked by 28-bp repeats, cloned into the pMK backbone using *SacI* and *EcoRI* restriction sites. (b) Schematic of each crRNA mapped onto the target sequence. The flanking PAM-like (red) or PAM (green) sequences are indicated and corresponding “core” or “seed” sequences are highlighted with a gray column. Each of the six spacers in CRISPR-1 and -2 target identical sequences that are flanked by a PAM-like motif, while CRISPR-3 spacers target proximal protospacers with a PAM. CRISPR-λ contains four spacer sequences that target four Lambda phage genes (J, O, R and E). CRISPR-DMS3 (negative control) contains seven identical spacers that target gene 24 in the DMS3 genome. (c) Sequences of the synthetic CRISPR loci. Spacers are separated by repeats (yellow).

CHAPTER THREE

DISTRIBUTION AND PHASING OF SEQUENCE MOTIFS THAT FACILITATE CRISPR ADAPTATION

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Andrew Santiago-Frangos

Contributions: Designed the research, performed the research, prepared manuscript figures, wrote the manuscript.

Co-Author: Murat Buyukyoruk

Contributions: Designed the research, prepared manuscript figures, performed the research.

Co-Author: Tanner Wiegand

Contributions: Designed the research, prepared manuscript figures, performed the research.

Co-Author: Pushya Krishna

Contributions: Designed the research, performed the research.

Co-Author: Blake Wiedenheft

Contributions: Designed the research, wrote the manuscript.

Manuscript Information

Andrew Santiago-Frangos, Murat Buyukyoruk, Tanner Wiegand, Pushya Krishna, Blake Wiedenheft

Current Biology

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- X Published in a peer-reviewed journal

Published by Elsevier, Inc.

Volume 31

Issue 16

23 August 2021

DOI: 10.1016/j.cub.2021.05.068

DISTRIBUTION AND PHASING OF SEQUENCE MOTIFS THAT
FACILITATE CRISPR ADAPTATION

Andrew Santiago-Frangos¹, Murat Buyukyoruk^{1,2}, Tanner Wiegand^{1,2}, Pushya Krishna^{1,2} and
Blake Wiedenheft¹

¹Department of Microbiology and Immunology, Montana State University, Bozeman, MT
59717, USA

²These authors contributed equally.

Abstract

CRISPR-associated proteins (Cas1 and Cas2) integrate foreign DNA at the “leader” end of CRISPR loci. Several CRISPR leader sequences are reported to contain a binding site for a DNA bending protein called Integration Host Factor (IHF). IHF-induced DNA bending kinks the leader of type I-E CRISPRs, recruiting an upstream sequence motif that helps dock Cas1-2 onto the first repeat of the CRISPR locus. To determine the prevalence of IHF-directed CRISPR adaptation, we analyzed 15,274 bacterial and archaeal CRISPR leaders. These experiments reveal multiple IHF binding sites and diverse upstream sequence motifs in a subset of the I-C, I-E, I-F and II-C CRISPR leaders. We identify subtype-specific motifs and show that the phase of these motifs is critical for CRISPR adaptation. Collectively, this work clarifies the prevalence and mechanism(s) of IHF-dependent CRISPR adaptation and suggests that leader sequences and adaptation proteins may coevolve under the selective pressures of foreign genetic elements like plasmids or phages.

Introduction

The repeat-spacer-repeat architecture characteristic of all CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) loci was first observed in 1987¹. However, the biological function of CRISPRs remained obscure until 2005, when three groups independently reported that CRISPR loci frequently contain “spacers” derived from foreign genetic elements²⁻⁴. By comparing CRISPR loci from closely related isolates of *Yersinia pestis*, Pourcel *et al.* noted that new phage-derived spacers are preferentially added to one end of the CRISPR⁴, which is often flanked by an adenine- and thymine-rich (AT-rich) sequence, previously termed the “leader”⁵. Collectively, these computational observations indicated that CRISPRs are part of an adaptive immune system that maintains a chronological record of previously encountered foreign genetic parasites^{6,7}. A role for CRISPR loci and Cas proteins in adaptive immunity was first established by phage challenge experiments performed by Barrangou *et al.*⁸. Polarized integration is crucial in most systems, since spacers at the leader-end of the CRISPR have been shown to provide greater levels of immunity⁹.

Structural and biochemical experiments aimed at understanding the mechanism(s) of polarized adaptation have resulted in two models for preferential integration of new spacer at the “leader-end”⁷. In type II systems, the Cas1 proteins have been shown to recognize the leader-repeat junction, while type I-E and I-F systems have been shown to rely on a DNA bending protein called Integration Host Factor (IHF)⁹⁻¹⁵. In the type I-E CRISPR system of *Escherichia coli* (K12), IHF binds to a roughly 30-base pair sequence motif that begins 6-base pairs upstream of the first repeat¹⁰. IHF-binding kinks DNA in the leader, creating a horseshoe shaped structure that stabilizes the Cas1-2 integrase complex bound to the first repeat of the CRISPR locus¹¹. One of the Cas1 dimers is wedged into the “toe” of the DNA horseshoe, resulting in specific contacts between the

Cas1 and IHF proteins, as well as sequence specific interactions between one lobe of Cas1 and an upstream sequence motif^{11,16}.

We recently noticed that the IHF binding site in the leader sequence of the type I-F CRISPR locus from *Pseudomonas aeruginosa* (PA14) is 8-base pairs further from the leader-repeat junction than what was originally observed in the type I-E systems of *Escherichia coli* (**Figure 3.1**). We hypothesized that this seemingly subtle difference has important mechanistic implications for CRISPR adaptation. The helical structure of double stranded DNA (dsDNA), contains ~11-base pairs per 360-degree rotation^{17,18}. Thus, the addition of 8-base pairs not only shifts the IHF binding site ~27 Å away from the leader-repeat junction, which would preclude previously observed Cas1-IHF interactions, but this insertion also introduces a ~260-degree rotation of the upstream DNA. This rotation would position one tip of the “DNA horseshoe” on the opposite side of the Cas1-2 integration complex from what has been previously observed for the *E. coli* integration complex¹¹.

Here we analyze 14,095 bacterial and 1,179 archaeal CRISPR leaders for conserved sequence motifs. This analysis reveals discretely distributed IHF binding sites and upstream motifs in a subset of I-C, I-E, I-F and II-C CRISPR loci. These leaders frequently contain multiple IHF binding sites and subtype-specific upstream motifs. The sequence, spacing, and orientation between motifs vary within and between subtypes, resulting in the detection of approximately 20 distinct leader architectures (**Figure S3.1**). Differences in leaders between closely related strains frequently involve insertions or deletions (indels) of 10-12 bps, which preserves the phase of these motifs, suggesting that phase is more important than distance *per se*. Additionally, we use *in vitro* adaptation assays to test the importance of both the sequence and the phase of leader motifs on new spacer integration in the type I-F system from *P. aeruginosa*. Overall, our data suggest that

the mechanisms of polarized CRISPR adaptation are diverse, but all systems that rely on IHF are expected to be phase-dependent.

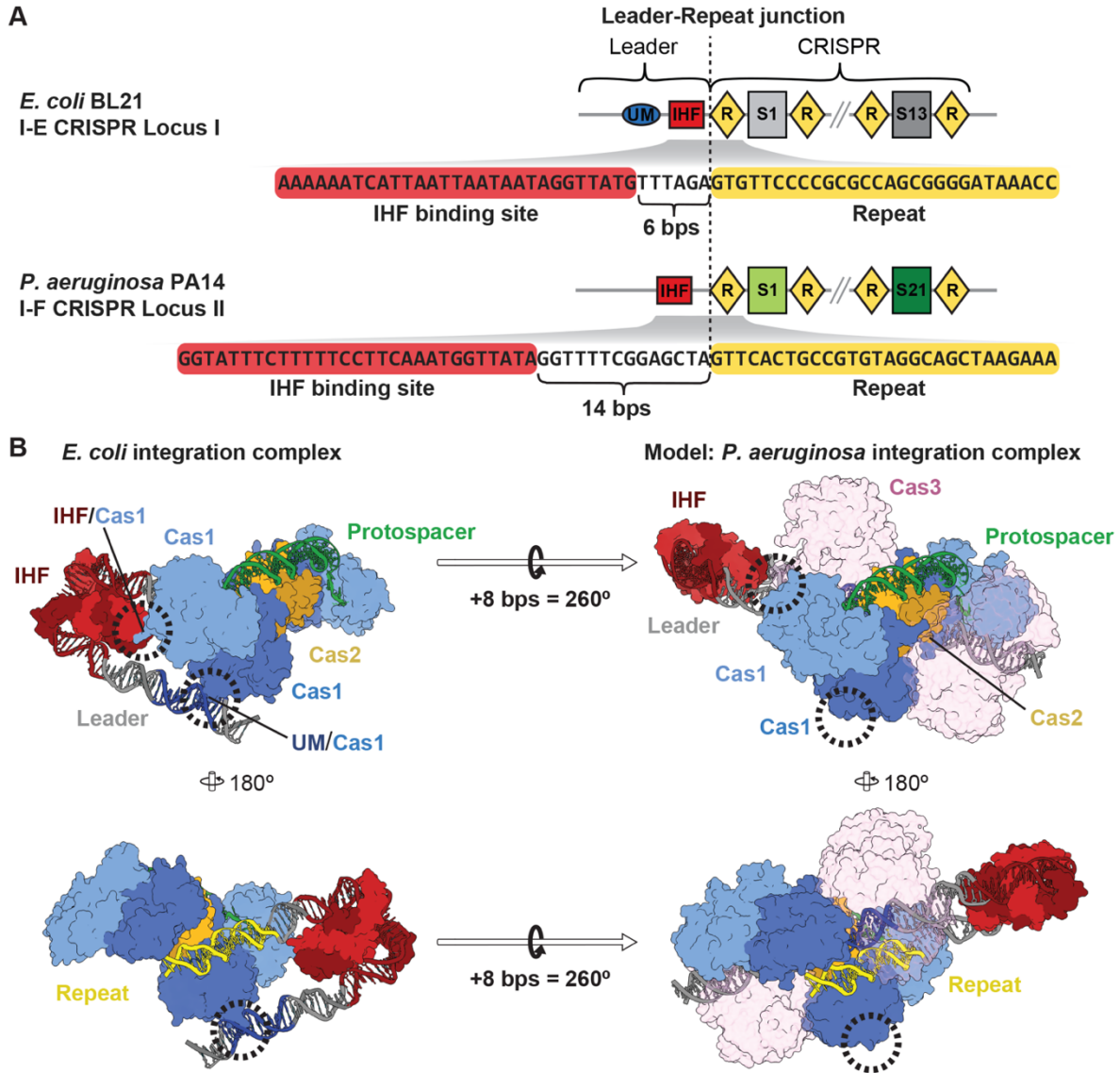


Figure 3.1. Spacing of sequence motifs in CRISPR leaders has mechanistic implications for the mechanism of spacer integration.

(A) Schemes of two CRISPR loci from *E. coli* BL21 (type I-E, top) and *P. aeruginosa* PA14 (type I-F, bottom), and sequences of the first CRISPR repeat and the upstream IHF binding site in the leaders. Repeating DNA motifs (diamonds) in the CRISPR locus are interspersed with unique spacer sequences (rectangles). The IHF binding site in the I-F system from *P. aeruginosa* PA14 is 8 bps further away from the leader-repeat junction, as compared to the I-E CRISPR locus. (B)

Structure of the previously determined type I-E integration intermediate (PDB: 5WFE). The Cas1-2 heterohexamer (blue and yellow) is in the process of integrating a protospacer (green) at the first CRISPR repeat (yellow). IHF kinking of DNA at the IHF binding site recruits an upstream motif (UM, dark blue) to interact with one lobe of a Cas1 subunit, while the second Cas1 subunit in the same dimer makes protein-protein contacts with IHF¹¹. Models of the 8 bp insertion present in *P. aeruginosa* CRISPR locus II illustrate how this insertion is predicted to disrupt both of these interactions by shifting IHF 27Å away from the integration complex and rotating the upstream motif (UM) 260°. Cas3 domains of the I-F Cas2/3 fusion are shown semi-transparently for clarity. See also Figure S1 and S3.

Results

Identification and Distribution of IHF and Upstream Motifs

IHF binding sites have been identified in the leader sequences of a few type I-E systems and in the type I-F system from *Pectobacterium atrosepticum*^{10,12,19}. To broadly determine the prevalence and distribution of IHF binding sites in CRISPR leaders, we queried all complete bacterial and archaeal genome sequences available at NCBI for CRISPR loci and associated leader sequences using CRISPRDetect²⁰. In total, we identified 15,274 CRISPR loci, representing most of the major subtypes (**Table 3.1**). We queried 200 base pairs upstream of each CRISPR using previously established position weight matrices for IHF binding sites^{21,22}. Sequences identified with this position weight matrix were used to construct more detailed position weight matrices, which were then used to iterate the search. In total, we identified 8,631 putative IHF binding sites in 15,274 leader sequences. Next, we culled the list of leaders by eliminating redundant sequences that were greater than 95% identical. This eliminates closely related genomes from a few organisms which were oversampled in our initial dataset (e.g., *E. coli*), and enables a more accurate representation of the phylogenetic distribution of IHF binding sites in leader sequences. In this subset of 6,533 non-redundant leaders, IHF binding sites were found in 24% of I-C (n=176), 26% of I-E (n=336), 83% of I-F (n=444), and 32% of II-C (n=108) leaders (**Table 3.1**). Interestingly,

most I-F leaders (53%) and several I-E, II-C and I-C leaders (6%, 6%, and 3%, respectively) contain more than one IHF binding site. We differentiate “proximal IHF” from “distal IHF” binding sites according to their position relative to the leader-repeat junction.

	Subtype	<i>Redundant</i>	Number of leaders		
			<i>Non-Redundant</i>		
		All	All	One or more IHF sites	Two or more IHF sites
Type I	I-A	296	187	27	3
	I-B	1516	761	238	36
	I-C	1324	754	178	24
	I-D	152	117	14	2
	I-E	5068	1329	335	84
	I-F	1683	536	444	279
	I-U	157	109	5	0
	I-V	2	2	0	0
Type II	II-A	823	178	39	4
	II-B	34	9	5	0
	II-C	662	337	103	17
Type III	III-A	853	242	93	11
	III-B	409	309	96	12
	III-C	29	23	6	0
	III-D	179	145	31	5
Type V	V-A	28	14	6	1
	V-B	3	3	0	0
Type VI	VI-A	2	2	1	0
	VI-B	56	16	9	2
	VI-C	2	1	0	0
	NA	1996	1459	315	32
Total		15274	6533	1945	512

Table 3.1. Subtype distribution of microbial CRISPRs and identified IHF binding sites.

15,274 CRISPR loci were identified in bacterial and archaeal genomes. Closely related microbial genomes, which causes an over-representation of certain CRISPR subtypes (I-B, I-C, I-E, I-F, II-A, II-C, III-A) (“Redundant” column), were systematically removed (“Non-Redundant” columns). The number of non-redundant CRISPR leaders in containing at least one, or at least two unique IHF binding sites is reported. See also Figure S3.1.

According to the previously established model for type I adaptation, the DNA bending behavior of IHF presents an upstream sequence motif that is recognized by specific amino acids

on Cas1 (i.e., R131 and R132)¹¹. In support of this model, the proximal IHF binding sites in I-E leaders are tightly distributed around a midpoint of 20-base pairs from the leader-repeat junction, and I-C and II-C leaders possess similar IHF sites at midpoints of 22 and 21-base pairs from the leader-repeat junction (**Figure S3.1**). However, the additional 8-base pairs that separate the IHF binding site from the leader-repeat junction in *P. aeruginosa* PA14, as compared to *E. coli* BL21, is preserved in a comparison of all type I-E and I-F leaders that contain IHF binding sites (**Figure S3.1**). This ~8-base pair insertion is unique to the type I-F systems and may be an evolutionary adaptation necessary to accommodate the Cas3 domain, which is uniquely fused to the Cas2 protein in I-F CRISPR system (**Figure 3.1**)^{23–25}. In addition to making room for Cas3, this insertion also rotates the upstream DNA by ~260-degree. This rotation suggests that either the Cas1-2/3 integration complex relies on a distinct mechanism for recognizing an upstream sequence motif, or that I-F systems do not rely on upstream sequence motifs. To distinguish between these two possibilities and to identify conserved upstream sequence motifs in other IHF containing leaders, we performed a *de novo* motif search using a combination of MEME and FIMO^{22,26}. MEME identified the upstream motif previously reported for type I-E systems^{11,19}, as well as several new motifs (**Figure 3.2**). The sequence and location of these motifs are characteristic of specific subtypes. In addition to the upstream motifs and proximal IHF binding sites, MEME also identified additional upstream IHF binding sites and additional, distally located upstream motifs that are either direct repeats or inverted repeats of the proximal upstream motif (**Figure 3.2**). The spacing between these motifs sometimes vary between leaders, but changes in the spacing are almost always restricted to increments of 10-12 base pairs, indicating that phase is conserved (**Figure**

S3.1). The I-C distal direct repeat and inverted repeat of the upstream motif are exceptions to this rule, which are often shifted by 8-base pairs (**Figure S3.1**).

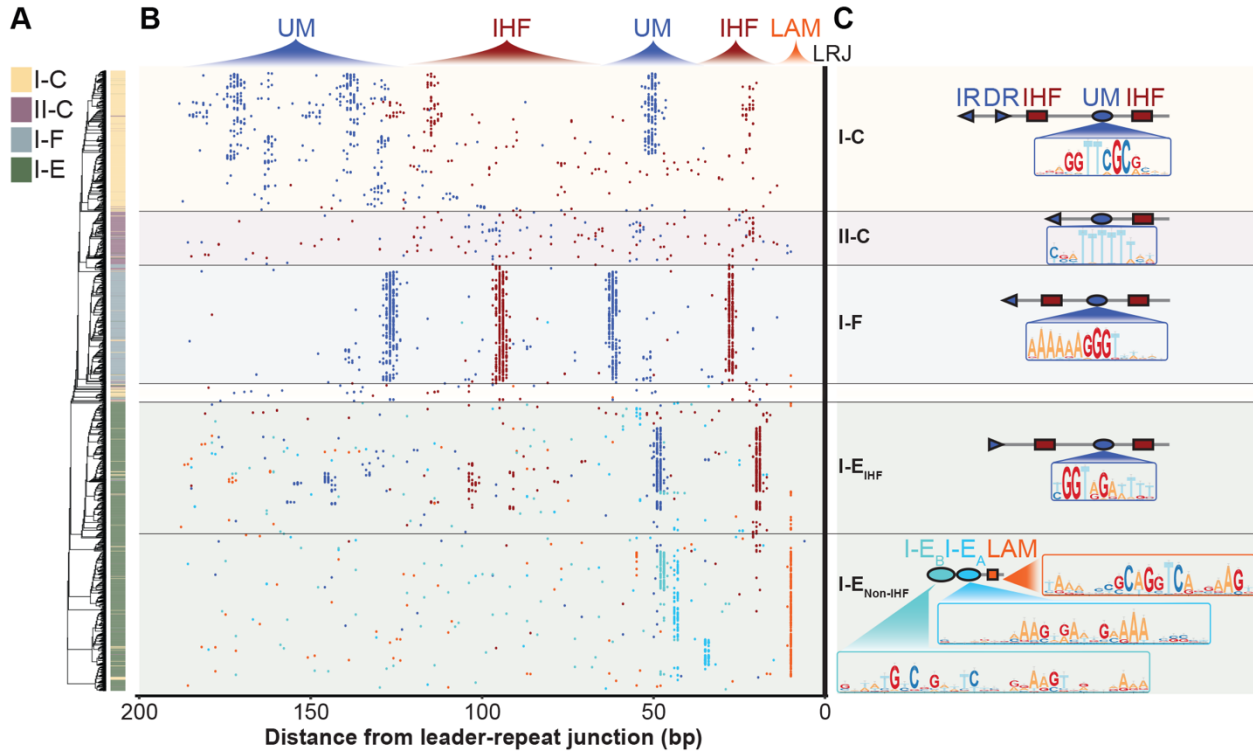


Figure 3.2. IHF-directed CRISPR adaptation is widespread.

(A) Phylogenetic tree generated from alignment of 200 bp of I-C, II-C, I-F and I-E CRISPR leaders and the first repeat. (B) Distributions of motifs within I-C, II-C, I-F and I-E CRISPR leaders. Each dot represents the midpoint of IHF binding sites (red), subtype specific upstream motifs (UMs) (dark blue), I-E_A (cyan), or I-E_B (teal), Leader Anchoring Motif (orange). Many of the leaders shown possess proximal IHF and UMs found between 0-70 bp upstream of the leader-repeat junction (LRJ), and distal IHF and UMs found 70-200 bp upstream of the LRJ. (C) Schematic of the prominent architecture for motifs within I-C, II-C, I-F and I-E CRISPR leaders. Position weight matrices are shown for subtype specific upstream motifs, and motifs found in I-E leaders that do not contain IHF binding sites. See also Figure S3.1.

While IHF has been shown to be critical for polarized CRISPR adaptation in the type I-E system from *E. coli*^{10,11}, we found that roughly 75% of I-E leaders do not contain a canonical IHF binding site. 20% of I-E leaders instead contain a “leader-anchoring motif” that directly abuts the

first repeat, in a similar manner to the “leader-anchoring site” reported for several of the type II systems (**Figure 3.2**)^{9,13–15}. While proximity of I-E leader-anchoring motifs to the leader-repeat junction is similar to those in type II systems, the sequences themselves are notably different. The type I-E leader-anchoring motifs contain an off-center TCA sequence, and a 5’ to 3’ TTR triplet on the opposite strand, which are hallmarks of IHF and Hbb DNA binding sites (**Figure S3.2**)^{27–31}, though these motifs are not detectable using any of the previously established position weight matrices for IHF binding sites (**Figure 3.2**). The IHF-like leader-anchoring motif may associate with DNA bending proteins that recruit upstream motifs. In fact, MEME identifies two motifs (I-E_A and I-E_B) that are unique to I-E leaders that contain “IHF-like leader-anchoring motifs” (**Figure 3.2**). If the “IHF-like leader-anchoring motif” is in fact a DNA bending sequence that functions to recruit the unique upstream motifs (i.e., I-E_A or I-E_B), then the proximity of this motif relative to the leader-repeat junction, would change the phase of the upstream motif relative to what has been observed for the IHF containing leader from *E. coli*. Thus, we hypothesized that Cas1 proteins associated with these systems may not rely on arginines 131 and 132 (R131 and R132), which are critical for upstream motif recognition by Cas1 in IHF-dependent CRISPRs from *E. coli*¹¹. To test this hypothesis, we aligned 368 Cas1 sequences from CRISPR systems that contain “IHF-like leader-anchoring motifs”. Unlike I-E Cas1 proteins that are associated with IHF containing leaders, these Cas1 proteins do not maintain R131 and R132, suggesting that the mechanism of adaptation in these CRISPRs is distinct (**Table 3.2**). These findings suggest these two types of I-E CRISPR systems may be phylogenetically distinct, indeed IHF-regulated I-E CRISPRs are found in bacteria belonging to the Phylum *Proteobacteria*, while leader-anchoring motif-regulated I-E CRISPRs are found in *Actinobacteria*. Collectively, these results reveal diverse upstream architectures that

include motifs anticipated to interact with DNA bending proteins (e.g., IHF, Hbb, or others) and DNA bending is anticipated to recruit upstream sequences that are critical for polarized adaptation in approximately 56% of all 15,274 CRISPR loci that we analyzed.

Subtype	Leader motifs	Number of Cas1s	R131		R132	
			<i>Identity</i>	<i>Similarity</i>	<i>Identity</i>	<i>Similarity</i>
I-E	IHF or UM	454	57.5%	68%	71.6%	72.9%
I-E	LAM, I-E _A , I-E _B	368	3.7%	5.4%	11.2%	13.8%
I-F	IHF or UM	499	11%	11%	11.9%	22.4%

Table 3.2. Cas1 residues that interact with canonical I-E UM are only conserved in Cas1s associated with IHF-regulated I-E CRISPRs.

IHF-binding Sites and Upstream Motif are Critical for Efficient Integration

To determine if the IHF binding sites and upstream motifs play a direct role in I-F CRISPR adaptation, we performed *in vitro* integration assays using the leader sequence derived from the CRISPR2 locus of *P. aeruginosa* PA14. According to our bioinformatic analysis, this leader contains both proximal and distal IHF binding sites, as well as proximal and distal upstream motifs (**Figure 3.2 and 3.3**). To determine which of these motifs participate in new spacer integration, we compared integration efficiencies measured for the wildtype leader sequence, to leaders where we either replaced the proximal IHF binding site with an IHF consensus sequence from *E. coli* (Opt. IHF prox), mutated key positions in the proximal IHF binding site (Mut. IHF prox), scrambled the proximal upstream motif (Mut. UM prox), deleted the distal IHF site (Del. IHF distal) or scrambled the distal upstream motif (Mut. UM distal) (**Figure 3.3C**).

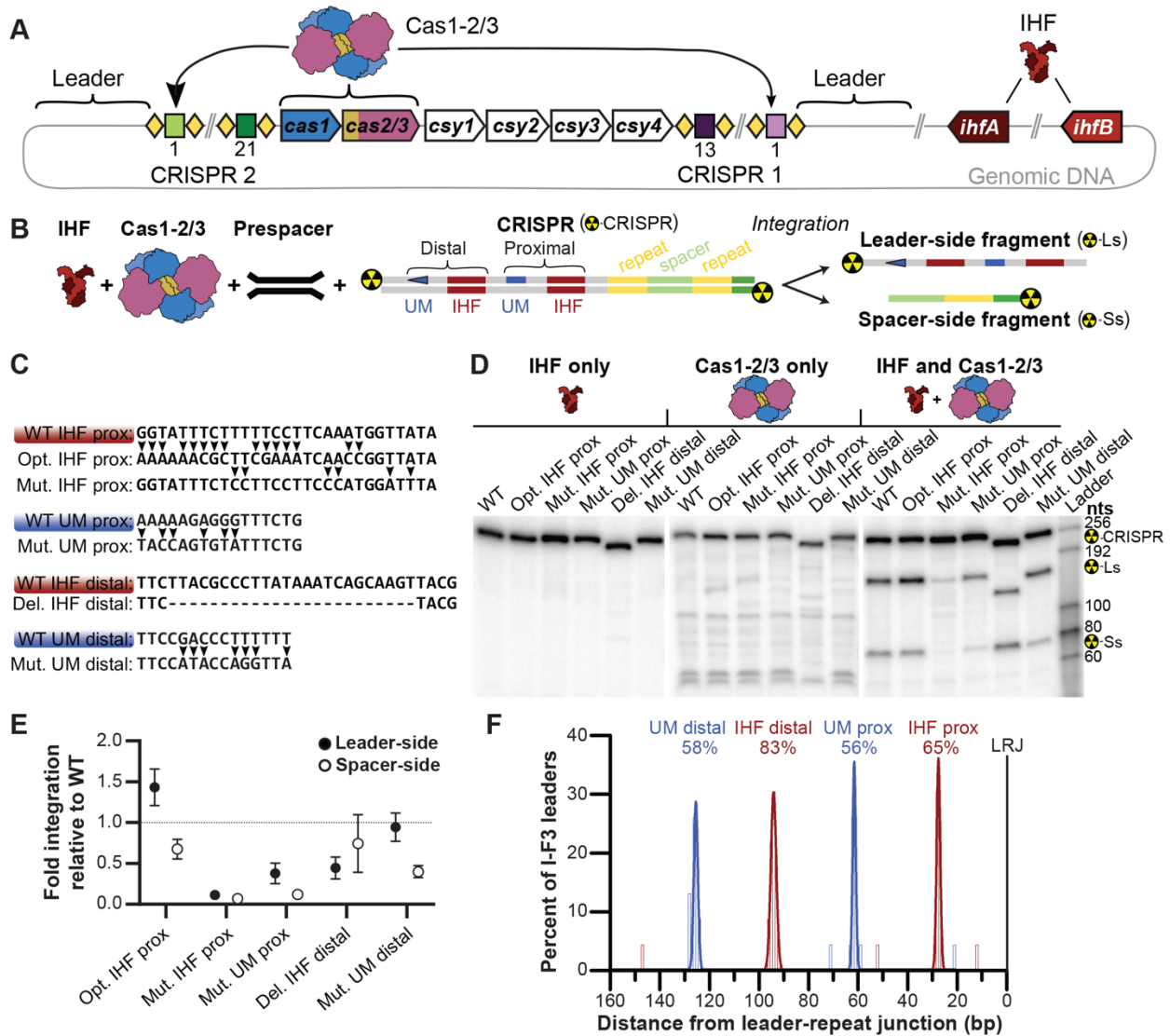


Figure 3.3. IHF binding sites and upstream motif are critical for efficient integration.

(A) Scheme of I-F CRISPR system from *P. aeruginosa* PA14. Two CRISPR loci composed of repeats (diamonds) and spacers, flank six *cas* genes (arrows). Cas1-2/3 assembles into a heterohexameric complex (blue, yellow, and purple, respectively) that catalyzes the integration of new spacers into CRISPR loci. IHF is a heterodimer of two related proteins encoded by *ihfA* (brown) and *ihfB* (tan). (B) Scheme for *in vitro* integration of a prespacer DNA into ^{32}P -labelled DNA derived from the CRISPR2 locus of *P. aeruginosa* PA14. Leader- and spacer-side transesterification reactions produce a large and a small ^{32}P -labelled DNA fragment respectively, which are separated by denaturing gel electrophoresis. Proximal and distal IHF binding sites and upstream motifs (UMs) in the leader are annotated. (C) IHF binding sites (brown) or UMs (blue) are shown. Two variants (arrows) of the proximal IHF site were tested (i.e. Opt IHF prox and Mut IHF prox.), one variant of the proximal UM (Mut. UM prox.), one variant of the distal IHF (Del. IHF idtal) and one variant of the distal UM (Mut. UM distal) were tested. (D) Representative

images of integration assays resolved on denaturing gels. Reactions were performed either in the presence of IHF alone (left), Cas1-2/3 alone (middle), or in the presence of both protein complexes (right). The expected positions of full-length CRISPR substrate (Del. IHF distal variant, 233 nts; WT and other variants, 255 nts), as well as leader side (Ls) (Del. IHF distal variant, 127 nts; WT and other variants, 151 nts) and spacer side (Ss) (all variants, 77 nts) integration products are indicated. **(E)** Quantification of leader- (dark gray) or spacer-side (light gray) integration. The average ($\pm$ s.d.) of triplicate reactions is shown. Integration was confirmed via high-throughput sequencing (**Figure S3.3**). Uncropped images and replicate gels are provided (**Figure S3.4**). **(F)** Tally of the midpoints of IHF binding site (red) and I-F upstream motifs (blue) identified in 23 non-redundant I-F3 CRISPR leaders³⁶. The area of histograms fit to peaks of motif matches are reported to convey the percentage of leaders with a given motif. Most I-F3 (Tn7-associated) CRISPR leaders possess motifs needed for efficient spacer integration. See also Figure S3.3-S3.5.

IHF-binding sites in the *P. aeruginosa* CRISPR2 leader deviate from the *E. coli* IHF consensus sequence at the 5' "A-tract" and central "WATCAR" regions, despite *P. aeruginosa* and *E. coli* IHF proteins sharing greater than 70% sequence identity (**Figure S3.2**)²⁷⁻²⁹. To determine whether an "optimized" IHF site would result in more efficient spacer integration, we replaced the IHF binding site in the I-F leader with the consensus IHF binding sequence from *E. coli*. These two sequences differ at 15 positions over 29 bps. Although optimization of the IHF binding site (*Opt. IHF prox*) resulted in slightly increased leader-side integration (1.4-fold), there is a concomitant decrease in spacer-side integration (1.5-fold) (**Figure 3.3D, 3.3E**). To verify that the IHF binding sequence is important for integration, we mutated six bases that are critical for IHF recognition (**Figure 3.3C**)^{27,28,32}. These mutations (*Mut. IHF prox*) result in large decreases for both leader- and spacer-side integration (8.7 and 14.7-fold respectively) (**Figure 3.3D, 3.3E, Figure S3.3 and S3.4**), which supports previous work from the Fineran lab demonstrating that the IHF protein is necessary for adaptation in the I-F system from *P. atrosepticum*¹². While it is expected that IHF is involved in DNA bending, there have been no prior reports of an upstream motif in I-F CRISPR leaders. To determine if the proximal upstream motif sequence we identified is important for CRISPR adaptation, we scrambled the sequence (*Mut. UM prox*). This mutation

results in a large decrease in both leader- and spacer-side integration efficiency (2.6- and 8.3-fold respectively) (**Figure 3.3C, 3.3D, 3.3E, Figure S3.3 and S3.4**).

Fagerlund *et al* previously detected several IHF-like sequences upstream of the I-F CRISPR locus in *P. atrosepticum*¹², but the importance of these distal sequences has gone untested. We hypothesized the distal IHF binding site and the distal upstream motif, which both occur at high frequency, would be functionally important for CRISPR adaptation. To test this hypothesis, we deleted the distal IHF site (*Del. IHF distal*) or scrambled the distal upstream motif (*Mut. UM distal*). Deletion of the distal IHF site reduces leader-side integration by 2.2-fold while changes to spacer side integration are within experimental error (**Figure 3.3D and 3.3E**). In contrast, scrambling the distal upstream motif sequence does not impact leader-side integration but reduces spacer-side integration by 2.5-fold. Collectively, these results suggest that the distal IHF and upstream motif sequences participate in new spacer integration.

Tn7-associated I-F3 Leaders Contain Motifs Needed for Efficient Integration

Recently, Petassi *et al* reported that Tn7-associated I-F3 CRISPR systems are accompanied by short CRISPRs³³. The spacers in I-F3 CRISPRs have recently been shown to guide transposition in a sequence-specific manner^{34,35}. To determine whether I-F3 CRISPR loci may continue to acquire new spacers that would enable the Tn7-like system to adapt to transpose into novel locations, we analyzed a subset of I-F3 leaders using custom position weight matrices for IHF binding sites and I-F upstream motifs developed above. In 23 non-redundant I-F3 leaders, 65% and 85% possess proximal and distal IHF binding sites (respectively), and 56-58% possess proximal and distal I-F upstream motifs (**Figure 3.3F**). Further, all four leader motifs are found in the same positions as in canonical I-F leaders (**Figure 3.2**). Since most I-F3 systems are not

associated with operons that encode *cas1* or *cas2/3*, it has been hypothesized the integration complex may be provided *in trans*³⁶. The conservation of IHF binding sites and upstream motifs in a subset of the I-F3 leaders, suggests that trans-acting Cas1-2/3 integration complexes may be compatible with these leaders. One barrier to the spread of mobile genetic elements is compatibility with host factors. Our *in vitro* integration results show that *P. aeruginosa* IHF can recognize an *E. coli* consensus IHF binding site (**Figure 3.3C, 3.3D, 3.3E**), corroborating previous reports that *E. coli* IHF may complement the integration of new spacers into I-F CRISPR loci by Cas1-2/3^{37,38}. Further, these observations suggest that I-C, II-C, I-E, I-F and Tn7-associated I-F3 CRISPR systems may exchange efficiently between microbes that encode IHF.

Phased Leader Motifs Facilitate CRISPR Adaptation

The insertion of 1, 2, or 5-base pairs between the leader-repeat junction and the IHF binding site was previously shown to progressively inhibit new spacer integration into a I-E CRISPR locus¹¹. These results were used to highlight the importance of the IHF-Cas1 interaction, which we have no reason to question. However, we hypothesized that these mutations would also alter the position of the upstream motif and that phase of the upstream motif might be more important than distance *per se*. To test this hypothesis, we measured the integration efficiency of new spacers into a fragment of a CRISPR from strain 14 of *P. aeruginosa* (PA14). Variants of the PA14 leader were designed to either preserve (e.g., +10 bp) or disrupt (e.g., +5 bp) phasing between either the leader-repeat junction and the proximal IHF, or the proximal IHF site and the proximal upstream motif, or both (**Figure 3.4A**).

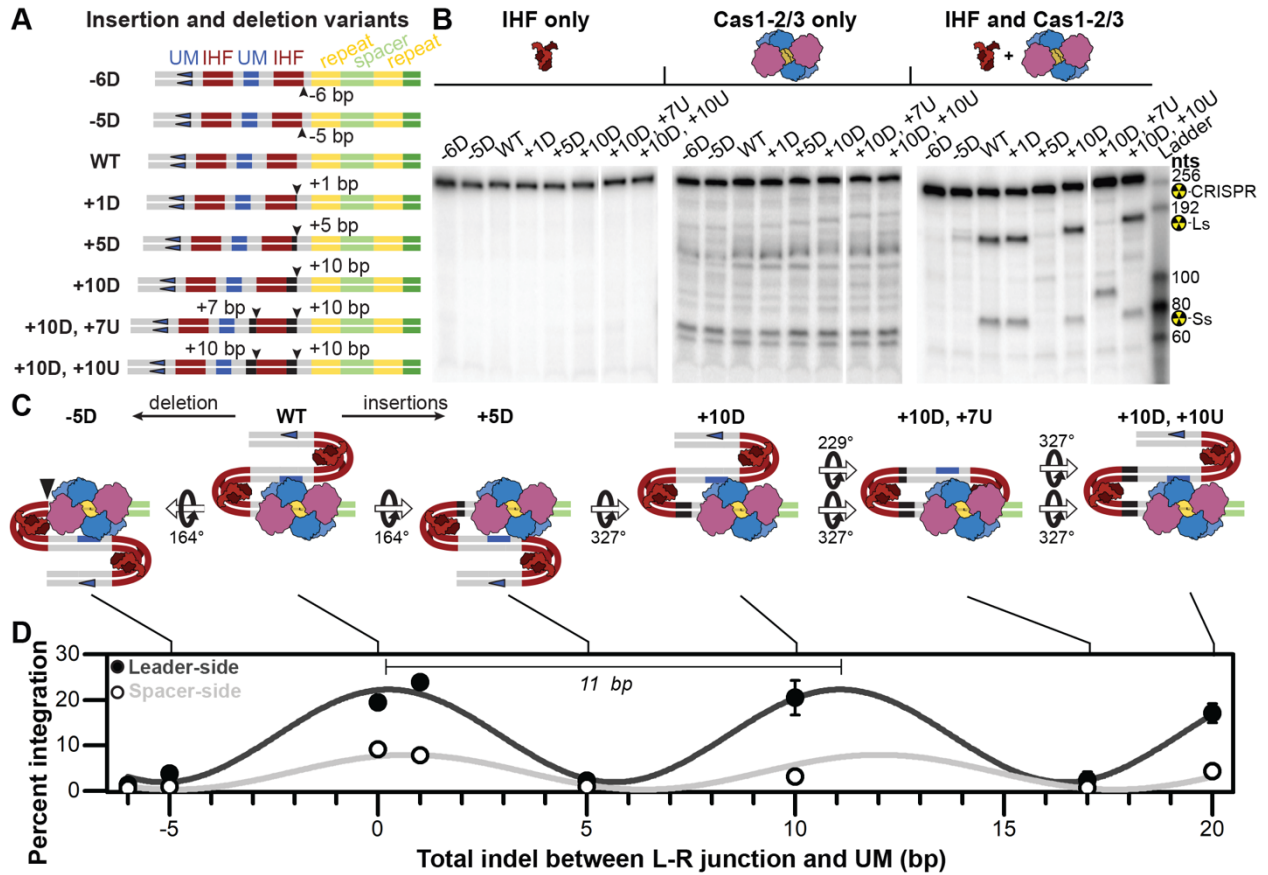


Figure 3.4. Phase-dependent CRISPR adaptation.

(A) Schematic of insertion and deletion (indel) variants generated to test the impact of distance and phase of leader motifs relative to the CRISPR locus. (B) I-F CRISPR integration is insensitive to indels that maintain phase, (10-base pairs), while indels of less than a full helical turn (five to seven base pairs) result in integration defects. Representative images of integration assays resolved on denaturing gels. Reactions were performed either in the presence of IHF alone (left), Cas1-2/3 alone (middle), or in the presence of both protein complexes (right). The expected positions of full-length substrate (249-275 nts depending on indel), as well as leader-side (145-161 nts depending on indel) and spacer-side (77 nts for all variants) integration products are indicated. (C) Schemes of putative IHF-bound conformations of wildtype and mutant leaders. Estimated rotations of DNA, relative to wildtype, are indicated. (D) Quantification of prespacer integration into all leader indel variants, either at leader- (dark gray) or spacer-side (light gray). The average ($\pm$ s.d.) of triplicate reactions is shown. Leader- and spacer-side integration data were independently fit to sine waves ($R^2=0.95$ and $R^2=0.75$), of wavelengths 10.8 ± 0.4 and 11.3 ± 0.7 bp, respectively. Integration was confirmed via high-throughput sequencing (Figure S3.3). Uncropped images and replicate gels are provided (Figure S3.6). See also Figure S3.3 and S3.6.

In the absence of IHF, the Cas1-2/3 proteins from *P. aeruginosa* inefficiently integrate prespacers into multiple sites along the leader (**Figure 3.3D, 3.4B**). In contrast, the addition of IHF to these reactions facilitates Cas1-2/3-mediated integration of prespacers at the leader- and spacer-side of the first repeat (**Figure 3.3D, 3.4B**). High-throughput sequencing of the integration reactions performed using primers at either end of the CRISPR DNA confirm that IHF decreases off-target spacer integration from 75.2% to 9.0% (**Figure S3.3**). Additionally, we observed that leader-side integration is more efficient than spacer-side integration and that 49.8% of unambiguously integrated prespacer substrates are trimmed prior to integration (**Figure S3.3**; see methods for explanation of ambiguous spacer integration events). These results are similar to those previously reported for *E. coli* I-E and *P. atrosepticum* I-F systems^{12,16,39}.

To specifically test the importance of phase, we designed a series of I-F leader mutants that maintain the motif, but perturb the phase of the proximal IHF and proximal upstream motifs to varying degrees (-6D, -5D, +1D, +5D) (**Figure 3.4A**). Consistent with our hypothesis, indels between the proximal IHF site and the leader-repeat junction, or the proximal IHF site and proximal upstream motif, both inhibit new spacer integration. The degree to which these changes impact the efficiency of integration correlates with changes to the phase. In other words, the more out of phase, the lower the efficiency of integration. Importantly, insertions that restore the original phase (i.e., +10D, +10U), also restore integration activity (**Figure 3.4**). Collectively, the data show that integration efficiencies oscillate with a wavelength of ~11 bp, corresponding to a complete turn of double stranded DNA helix (**Figure 3.4D**)^{17,18}. These experiments indicate that the phase of leader motifs, rather than their distance from the leader-repeat junction, is critical for efficient integration of new spacers at the leader end of the CRISPR locus.

Discussion

Since the discovery of regulatory elements in DNA⁴⁰, there has been an interest in determining the “grammatical rules” of sequence motifs that regulate the storage and retrieval of genetic information. The regulatory influence of a sequence motif is determined by its location and orientation to other *cis* or *trans* acting sequence motifs, which often assemble into structures that regulate the activity of neighboring genetic elements^{41–43}.

Specific and efficient integration of new spacers into CRISPR loci is the crucial first step in the adaptive immune response of bacteria and archaea. Here we identify leader sequences and structural elements that regulate CRISPR adaptation. These “grammatical rules” help explain the mechanisms by which CRISPRs evolve and may be important for the design of CRISPR-based data recording applications^{44,45}. Moreover, we anticipate that motifs identified here may be used to improve computational methods designed to identify and assign CRISPR loci to particular subtypes.

While new spacers are generally added to one end of the CRISPR, there are several known examples of non-random integration of spacers outside the CRISPR locus^{10,12,14,46}. To determine whether off-target integration observed in *E. coli* could be explained by the presence of specific motifs, we queried 200 bps flanking both sides of 697 integration sites to identify I-E repeats, IHF binding sites, and upstream motifs using FIMO^{22,46}. Only 73 of these sites possess a significant match to a I-E repeat, 31 possess an IHF-like binding site, and 44 possessed a canonical I-E upstream motif. Importantly, none of these sites contain all three motifs. While the mechanistic basis for integration at these non-CRISPR sites remains unclear, the overexpression of Cas1 and Cas2 may explain some of the promiscuity reported in these experiments⁴⁶.

Previous studies have demonstrated that some type I CRISPRs rely on leader proximal and distal motifs^{11,12,16,19,47–49}. Conversely, in type II systems, Cas1-2 directly recognizes a leader anchoring motif located directly adjacent to the first repeat^{9,13–15,50}. Our work suggests that the mechanisms for adaptation may be more complex and that some II-C CRISPRs may require IHF, while 20% of type I-E CRISPRs appear to rely on a leader anchoring motif that is unique in sequence but similar in position to what has been observed for the type II systems (**Figure 3.2**). Bioinformatic analyses suggest IHF is predominantly restricted to *Proteobacteria*, although similar type II DNA-binding proteins may be found in other microbes^{30,51}. Perhaps as expected, CRISPR leaders in which we identified phased IHF binding sites and upstream motifs (I-E, I-F, I-C and II-C) are highly represented in *Proteobacteria*⁵². Further, the split between I-E leaders that contain IHF sites versus those that contain leader-anchoring motifs roughly coincides with a split between I-E CRISPRs originating from *Proteobacteria* versus those originating from *Actinobacteria*. These observations point to a continuing co-evolution of CRISPR adaptation complexes with host machinery to support the fidelity of new spacer integration.

Leader sequences are diverse and variations in leader architectures may tune adaptation rates. In fact, many microbes possess multiple CRISPR loci, which evolve at different rates^{52–54}. For example, I-F *cas* genes in *P. aeruginosa* are flanked by a short CRISPR locus (CRISPR1) that evolves slowly and a longer CRISPR locus (CRISPR2) that has been shown to evolve more rapidly (**Figure S3.5**)^{54,55}. *P. atrosepticum* SCRI1043 similarly possesses three CRISPR loci that evolve at different rates⁵³. In corroboration with the biochemical data presented here, the fastest evolving I-F CRISPR loci in *P. aeruginosa* PA14 and *P. atrosepticum* SCRI1043 contain leaders with proximal and distal IHF binding sites, as well as upstream motifs. Whereas slower evolving

CRISPR loci in these bacteria lack the proximal IHF binding site in their respective leaders (**Figure S3.5**). Collectively, these data suggest that CRISPR loci with leaders containing the full complement of motifs may adapt rapidly. Although rapidly evolving CRISPR loci may enable a quick response to a new pathogen, the dsDNA breaks associated with the integration of a new spacer may cause such loci to be prone to lose previously acquired spacers via homologous recombination between repeats or other DNA repair pathways^{56,57}. Thus, there may be a selective pressure that maintains CRISPR loci that have been tuned for either fast or slow adaptation.

The work presented here started with the observation that there are 8 additional base pairs between the leader-repeat junction and the first repeat of the CRISPR in *P. aeruginosa* (PA14), as compared to the CRISPRs in *E. coli*. This 8 bp insertion is not unique to the CRISPR in *P. aeruginosa* (PA14) but is instead conserved in I-F systems (**Figure S3.1**). While we show that IHF binding sites are common in I-E and I-F systems they are certainly not the rule. In fact, 75% of non-redundant I-E leaders do not contain IHF binding sites. Rather, 20% of I-E systems contain a leader-anchoring motif and one or two different unique upstream sequence motifs (**Figure 3.2**). These observations led us to look more broadly for IHF binding sites and unique upstream motifs, revealing that many leaders contain multiple IHF binding sites and diverse upstream sequence motifs that are characteristic of the I-C, I-E, I-F and II-C leaders. The conservation of relative distances between motifs suggested that phase, rather than distance, might be critical for efficient adaption in IHF-dependent systems. We tested the importance of phase in the I-F system and demonstrate insertions that are in phase with the DNA helix, restore efficient adaption (**Figure 3.4**). While leaders are expected to be critical for adaption in most systems, most leaders are also expected to control CRISPR expression^{58–60}. Ultimately, we expect that the motifs identified here

are only the tip of the iceberg and that additional bioinformatic studies will provide ongoing insights that will lead to a more comprehensive understanding of how leader sequences integrate signals for expression and adaptation of CRISPRs.

Acknowledgements

We thank Metin Balaban and Dr. Niema Moshiri for help using TreeCluster, Dr. Christopher Brown for help with CRISPRDetect. Dr. Ewelina Malecka-Grajek for generous assistance with imaging. Jesse Thomas for maintenance of the phosphor screen imaging system. Dr. Dieter Jahn and Dr. Christian-Alexander Dudek for submission of sequences to PRODORIC. Thanks to Coltran Hopman-Nichols for computational support and members of the B.W. laboratory for critical feedback and thoughtful discussion. A.S-F. is a postdoctoral fellow of the Life Science Research Foundation that is supported by the Simons Foundation. A.S-F. is supported by the Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund. Research in the Wiedenheft lab is supported by the NIH (R35GM134867), the M.J. Murdock Charitable Trust, a young investigator award from Amgen, and the Montana State University Agricultural Experimental Station (USDA NIFA). P.K. is supported by the National Institute of General Medical Sciences of the NIH (P20GM103474). Funders had no role in designing, performing, interpreting, or submitting the work.

B.W. is the founder of SurGene, LLC, VIRIS Detection Systems Inc. and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. A.S-F. is a co-inventor on a patent for a CRISPR-based diagnostic.

References

1. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* [Internet]. 1987;169(12):5429–33. Available from: <https://jb.asm.org/content/169/12/5429>
2. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* [Internet]. 2005 Aug 1;151(8):2551–61. Available from: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.28048-0>
3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* [Internet]. 2005 Feb;60(2):174–82. Available from: <http://link.springer.com/10.1007/s00239-004-0046-3>
4. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* [Internet]. 2005 Mar 1;151(3):653–63. Available from: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.27437-0>
5. Jansen R, Embden JDA van, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* [Internet]. 2002 Mar;43(6):1565–75. Available from: <http://doi.wiley.com/10.1046/j.1365-2958.2002.02839.x>
6. Vale PF, Little TJ. CRISPR-mediated phage resistance and the ghost of coevolution past. *Proc R Soc B Biol Sci* [Internet]. 2010 Jul 22;277(1691):2097–103. Available from: <https://royalsocietypublishing.org/doi/10.1098/rspb.2010.0055>
7. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJJ. CRISPR-Cas: Adapting to change. *Science* (80-) [Internet]. 2017 Apr 7;356(6333):eaal5056. Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aal5056>
8. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* (80-) [Internet]. 2007 Mar 23;315(5819):1709–12. Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.1138140>
9. McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol Cell* [Internet]. 2016 Nov;64(3):616–23. Available from: <http://dx.doi.org/10.1016/j.molcel.2016.08.038>
10. Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell* [Internet]. 2016 Jun;62(6):824–33. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1097276516301034>

11. Wright A V., Liu J-J, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the CRISPR genome integration complex. *Science* (80-) [Internet]. 2017 Sep 15;357(6356):1113–8. Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aao0679>
12. Fagerlund RD, Wilkinson ME, Klykov O, Barendregt A, Pearce FG, Kieper SN, Maxwell HWR, Capolupo A, Heck AJR, Krause KL, et al. Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci* [Internet]. 2017 Jun 13 [cited 2018 Jan 3];114(26):201618421. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28611213>
13. Wei Y, Chesne MT, Terns RM, Terns MP. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res* [Internet]. 2015 Feb 18;43(3):1749–58. Available from: <http://academic.oup.com/nar/article/43/3/1749/2411552/Sequences-spanning-the-leaderrepeat-junction>
14. Wright A V., Doudna JA. Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol* [Internet]. 2016 Oct 5 [cited 2017 Dec 9];23(10):876–83. Available from: <http://www.nature.com/doifinder/10.1038/nsmb.3289>
15. Xiao Y, Ng S, Nam KH, Ke A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* [Internet]. 2017 Oct 4 [cited 2017 Nov 7];550(7674):137–41. Available from: <http://www.nature.com/doifinder/10.1038/nature24020>
16. Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell*. 2016;62(6):824–33.
17. Lee DH, Schleif RF. In vivo DNA loops in *araCBAD*: size limits and helical repeat. *Proc Natl Acad Sci* [Internet]. 1989 Jan 1;86(2):476–80. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.86.2.476>
18. Wang JC. Helical repeat of DNA in solution. *Proc Natl Acad Sci* [Internet]. 1979 Jan 1;76(1):200–3. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.76.1.200>
19. Yoganand KNR, Sivathanu R, Nimkar S, Anand B. Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res* [Internet]. 2017 Jan 9;45(1):367–81. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1151>
20. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* [Internet]. 2016 Dec 17;17(1):356. Available from: <http://dx.doi.org/10.1186/s12864-016-2627-0>
21. Eckweiler D, Dudek C-A, Hartlich J, Brötje D, Jahn D. PRODORIC2: the bacterial gene regulation database in 2018. *Nucleic Acids Res* [Internet]. 2018 Jan 4;46(D1):D320–6. Available from: <http://academic.oup.com/nar/article/46/D1/D320/4607805>

22. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* [Internet]. 2011 Apr 1;27(7):1017–8. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr064>
23. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat Rev Microbiol* [Internet]. 2015 Nov 28;13(11):722–36. Available from: <http://dx.doi.org/10.1038/nrmicro3569>
24. Richter C, Gristwood T, Clulow JS, Fineran PC. In Vivo Protein Interactions and Complex Formation in the *Pectobacterium atrosepticum* Subtype I-F CRISPR/Cas System. *PLoS One*. 2012;7(12).
25. Rollins MF, Chowdhury S, Carter J, Golden SM, Wilkinson RA, Bondy-Denomy J, Lander GC, Wiedenheft B. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci* [Internet]. 2017 Apr 24 [cited 2017 Dec 9];114(26):201616395. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28438998>
26. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* [Internet]. 2009 Jul 1;37(Web Server):W202–8. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp335>
27. Aeling KA, Opel ML, Steffen NR, Tretyachenko-Ladokhina V, Hatfield GW, Lathrop RH, Senear DF. Indirect Recognition in Sequence-specific DNA Binding by *Escherichia coli* Integration Host Factor. *J Biol Chem* [Internet]. 2006 Dec;281(51):39236–48. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0021925819337901>
28. Rice PA, Yang S, Mizuuchi K, Nash HA. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell* [Internet]. 1996 Dec;87(7):1295–306. Available from: [http://dx.doi.org/10.1016/S0092-8674\(00\)81824-3](http://dx.doi.org/10.1016/S0092-8674(00)81824-3)
29. Swinger KK, Rice PA. IHF and HU: flexible architects of bent DNA. *Curr Opin Struct Biol* [Internet]. 2004 Feb;14(1):28–35. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0959440X0300188X>
30. Mouw KW, Rice PA. Shaping the *Borrelia burgdorferi* genome: crystal structure and binding properties of the DNA-bending protein Hbb. *Mol Microbiol* [Internet]. 2007 Mar;63(5):1319–30. Available from: <http://doi.wiley.com/10.1111/j.1365-2958.2007.05586.x>
31. Kobryn K, Naigamwalla DZ, Chaconas G. Site-specific DNA binding and bending by the *Borrelia burgdorferi* Hbb protein. *Mol Microbiol* [Internet]. 2000 Jul;37(1):145–55. Available from: <http://doi.wiley.com/10.1046/j.1365-2958.2000.01981.x>
32. Connolly M, Arra A, Zvoda V, Steinbach PJ, Rice PA, Ansari A. Static Kinks or Flexible Hinges: Multiple Conformations of Bent DNA Bound to Integration Host Factor Revealed by

Fluorescence Lifetime Measurements. *J Phys Chem B* [Internet]. 2018 Dec 13;122(49):11519–34. Available from: <https://pubs.acs.org/doi/10.1021/acs.jpcc.8b07405>

33. Petassi MT, Hsieh S-C, Peters JE. Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons. *Cell* [Internet]. 2020 Dec;183(7):1757-1771.e18. Available from: <https://doi.org/10.1016/j.cell.2020.11.005>

34. Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* [Internet]. 2019 Jul 12 [cited 2019 Aug 6];571(7764):219–25. Available from: <http://www.nature.com/articles/s41586-019-1323-z>

35. Halpin-Healy TS, Klompe SE, Sternberg SH, Fernández IS. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature* [Internet]. 2020 Jan 18 [cited 2020 Jan 28];577(7789):271–4. Available from: <http://www.nature.com/articles/s41586-019-1849-0>

36. Peters JE, Makarova KS, Shmakov S, Koonin E V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci* [Internet]. 2017 Aug 29;114(35):E7358–66. Available from: <http://www.pnas.org/lookup/doi/10.1073/pnas.1709035114>

37. Vorontsova D, Datsenko KA, Medvedeva S, Bondy-Denomy J, Savitskaya EE, Pougach K, Logacheva M, Wiedenheft B, Davidson AR, Severinov K, et al. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. *Nucleic Acids Res* [Internet]. 2015 Dec 15;43(22):10848–60. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1261>

38. Wiegand T, Semenova E, Shiriaeva A, Fedorov I, Datsenko K, Severinov K, Wiedenheft B. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. *Cris J* [Internet]. 2020 Oct 1;3(5):378–87. Available from: <https://www.liebertpub.com/doi/10.1089/crispr.2020.0069>

39. Kim S, Loeff L, Colombo S, Jergic S, Brouns SJJ, Joo C. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* [Internet]. 2020;579(7797):141–5. Available from: <http://dx.doi.org/10.1038/s41586-020-2018-1>

40. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* [Internet]. 1961 Jun;3(3):318–56. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022283661800727>

41. Pérez-Martín J, Rojo F, de Lorenzo V. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol Rev* [Internet]. 1994;58(2):268–90. Available from: <https://mmbr.asm.org/content/58/2/268>

42. Schleif R. DNA Looping. *Annu Rev Biochem* [Internet]. 1992 Jun 8;61(1):199–223. Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.3353710>

43. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* [Internet]. 2014 Jul 10;15(7):453–68. Available from: <http://www.nature.com/articles/nrg3684>
44. Shipman SL, Nivala J, Macklis JD, Church GM. Molecular recordings by directed CRISPR spacer acquisition. *Science* (80-) [Internet]. 2016 Jul 29;353(6298):aaf1175. Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.aaf1175>
45. Schmidt F, Cherepkova MY, Platt RJ. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* [Internet]. 2018 Oct 3 [cited 2018 Oct 30];562(7727):380–5. Available from: <http://www.nature.com/articles/s41586-018-0569-1>
46. Nivala J, Shipman SL, Church GM. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat Microbiol* [Internet]. 2018 Mar 29;3(3):310–8. Available from: <http://dx.doi.org/10.1038/s41564-017-0097-z>
47. Kieper SN, Almendros C, Brouns SJJ. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiol Lett* [Internet]. 2019 Jun 1;366(11):2016–20. Available from: <https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnz129/5525085>
48. Rollie C, Graham S, Rouillon C, White MF. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res* [Internet]. 2018 Feb 16;46(3):1007–20.
49. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*. 2012;40(12):5569–76.
50. Kim JG, Garrett S, Wei Y, Graveley BR, Terns MP. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR–Cas system. *Nucleic Acids Res* [Internet]. 2019 Sep 19;47(16):8632–48. Available from: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz677/5545008>
51. Kamashev D, Agapova Y, Rastorguev S, Talyzina AA, Boyko KM, Korzhenevskiy DA, Vlaskina A, Vasilov R, Timofeev VI, Rakitina T V. Comparison of histone-like HU protein DNA-binding properties and HU/IHF protein sequence alignment. *PLoS One*. 2017;12(11):1–24.
52. Bernheim A, Bikard D, Touchon M, Rocha EPC. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic Acids Res* [Internet]. 2019 Nov 20;48(2):748–60. Available from: <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1091/5634034>
53. Richter C, Dy RL, McKenzie RE, Watson BNJ, Taylor C, Chang JT, McNeil MB, Staals RHJ, Fineran PC. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res*. 2014;42(13):8516–26.

54. Westra ER, van Houte S, Oyesiku-Blakemore S, Makin B, Broniewski JM, Best A, Bondy-Denomy J, Davidson A, Boots M, Buckling A. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr Biol* [Internet]. 2015 Apr;25(8):1043–9. Available from: <http://dx.doi.org/10.1016/j.cub.2015.01.065>
55. Heussler GE, Miller JL, Price CE, Collins AJ, O'Toole GA. Requirements for *Pseudomonas aeruginosa* type I-F CRISPR-Cas adaptation determined using a biofilm enrichment assay. *J Bacteriol*. 2016;198(22):3080–90.
56. Kupczok A, Landan G, Dagan T. The Contribution of Genetic Recombination to CRISPR Array Evolution. *Genome Biol Evol*. 2015;7(7):1925–39.
57. Gudbergssdottir S, Deng L, Chen Z, Jensen JVK, Jensen LR, She Q, Garrett RA. Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol Microbiol* [Internet]. 2011 Jan;79(1):35–49. Available from: <http://doi.wiley.com/10.1111/j.1365-2958.2010.07452.x>
58. Pul Ü, Wurm R, Arslan Z, Geißen R, Hofmann N, Wagner R. Identification and characterization of *E. coli* CRISPR- cas promoters and their silencing by H-NS. *Mol Microbiol* [Internet]. 2010 Mar;75(6):1495–512. Available from: <http://doi.wiley.com/10.1111/j.1365-2958.2010.07073.x>
59. Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, Glover CVC, Graveley BR, Terns RM, Terns MP. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *S. treptococcus thermophilus*. *Mol Microbiol* [Internet]. 2014 Jul;93(1):98–112. Available from: <http://doi.wiley.com/10.1111/mmi.12644>
60. Przybilski R, Richter C, Gristwood T, Clulow JS, Vercoe RB, Fineran PC. Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biol* [Internet]. 2011 May 27;8(3):517–28. Available from: <http://www.tandfonline.com/doi/abs/10.4161/rna.8.3.15190>
61. van Dijk M, Bonvin AMJJ. 3D-DART: A DNA structure modelling server. *Nucleic Acids Res*. 2009;37(SUPPL. 2):235–9.
62. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci*. 2018;27(1):14–25.
63. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70–82.
64. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings Int Conf Intell Syst Mol Biol* [Internet]. 1994;2:28–36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7584402>

65. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* [Internet]. 2004 Jun;14(6):1188–90. Available from: <ftp://ftp.ncbi.nih.gov/genomes/Bacteria>
66. Pourcel C, Touchon M, Villeriot N, Vernadet JP, Couvin D, Toffano-Nioche C, Vergnaud G. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res*. 2020;
67. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* [Internet]. 2006 Jul 1;22(13):1658–9. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158>
68. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* [Internet]. 2012 Dec;28(23):3150–2. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts565>
69. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* [Internet]. 2013 Apr 1;30(4):772–80. Available from: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst010>
70. Gouveia-Oliveira R, Sackett PW, Pedersen AG. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics* [Internet]. 2007 Dec 28;8(1):312. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-312>
71. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. *PLoS One* [Internet]. 2010 Mar 10;5(3):e9490. Available from: <https://dx.plos.org/10.1371/journal.pone.0009490>
72. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. <scp>ggtree</scp> : an <scp>r</scp> package for visualization and annotation of phylogenetic trees with their covariates and other associated data. McInerney G, editor. *Methods Ecol Evol* [Internet]. 2017 Jan 22;8(1):28–36. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628>
73. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Battistuzzi FU, editor. *Mol Biol Evol* [Internet]. 2018 Dec 1;35(12):3041–3. Available from: <https://academic.oup.com/mbe/article/35/12/3041/5142656>
74. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinforma* [Internet]. 2020 Mar 5;69(1). Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.96>
75. Wickham H. ggplot2 [Internet]. Cham: Springer International Publishing; 2016. (Use R!). Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19791908>

76. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *J Open Source Softw* [Internet]. 2019 Nov 21;4(43):1686. Available from: <https://joss.theoj.org/papers/10.21105/joss.01686>
77. Qiu D, Damron FH, Mima T, Schweizer HP, Yu HD. PBAD-Based Shuttle Vectors for Functional Analysis of Toxic and Highly Regulated Genes in *Pseudomonas* and *Burkholderia* spp. and Other Bacteria. *Appl Environ Microbiol* [Internet]. 2008 Dec 1;74(23):7422–6. Available from: <https://aem.asm.org/content/74/23/7422>
78. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* [Internet]. 2014 Mar 1;30(5):614–20. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt593>
79. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet]. 2009;10(1):421. Available from: <http://www.biomedcentral.com/1471-2105/10/421>

Supplemental Figures

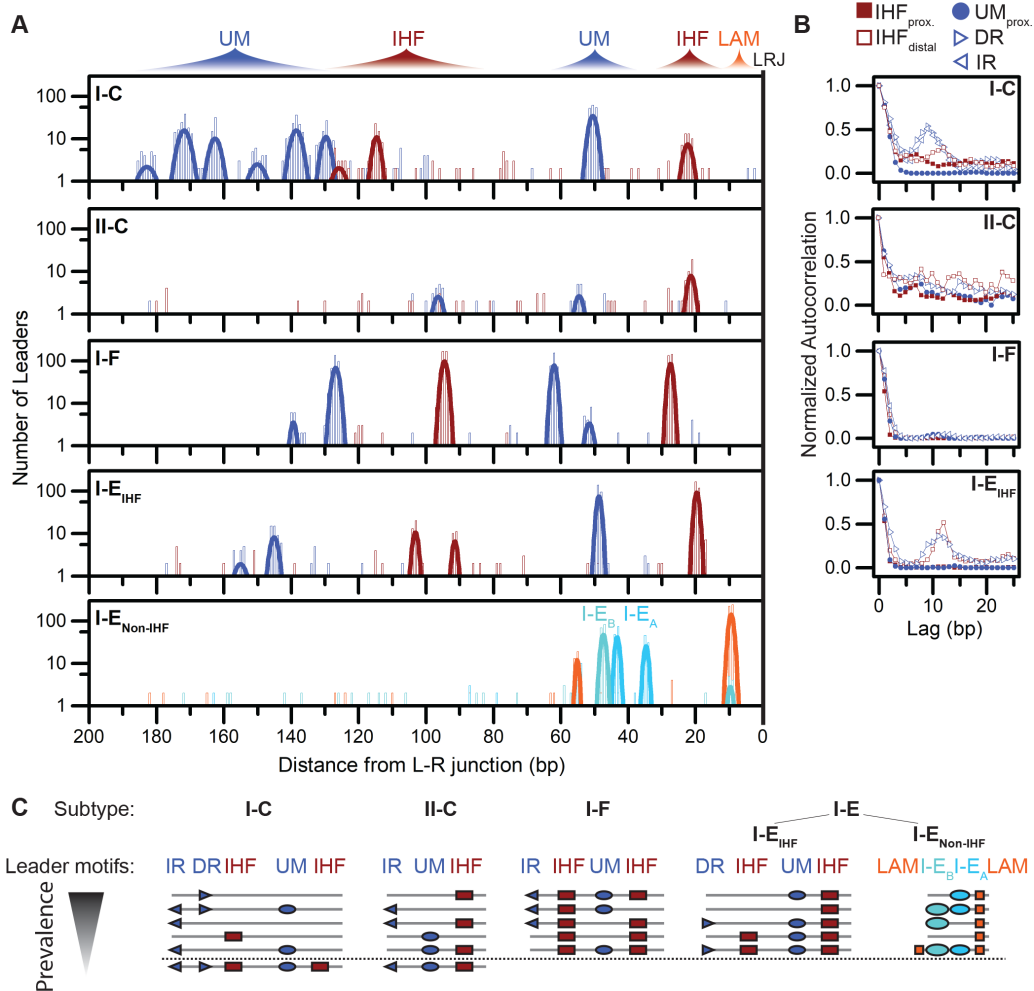


Figure S3.1. Naturally occurring indels preserve the phase of leader motifs, Related to Figures 3.2-3.4 and Table 3.1.

(A) Tally of the midpoints of motifs found in I-C, II-C, I-F, I-E_{IHF}, I-E_{Non-IHF} leaders. Phased motifs are fit to gaussian curves. (B) Autocorrelation plots of the midpoints of motifs found in I-C, II-C, I-F, I-E_{IHF}, I-E_{Non-IHF} leaders. Common indels appear as peaks in the autocorrelation function. Motifs that are frequently shifted by 10-11 bp relative to the LRJ are retained in phase with the DNA helical repeat. (C) Schematics of observed leader motif architectures within each CRISPR subtype. IHF binding sites (IHF) are shown as red rectangles, upstream motifs (UMs) are shown in as blue ovals. The UM motif is often repeated, as a direct repeat (DR) and/or an inverted repeat (IR). The sequences of the UMs and the relative distances between leader motifs varies between leader subtypes. Many I-F leaders possess all four motifs, whereas most I-E leaders possess only the proximal UM and IHF site. I-C and II-C leaders with the full complement of leader motifs are rare, and not among the top five most prevalent leader architectures (below dotted line).

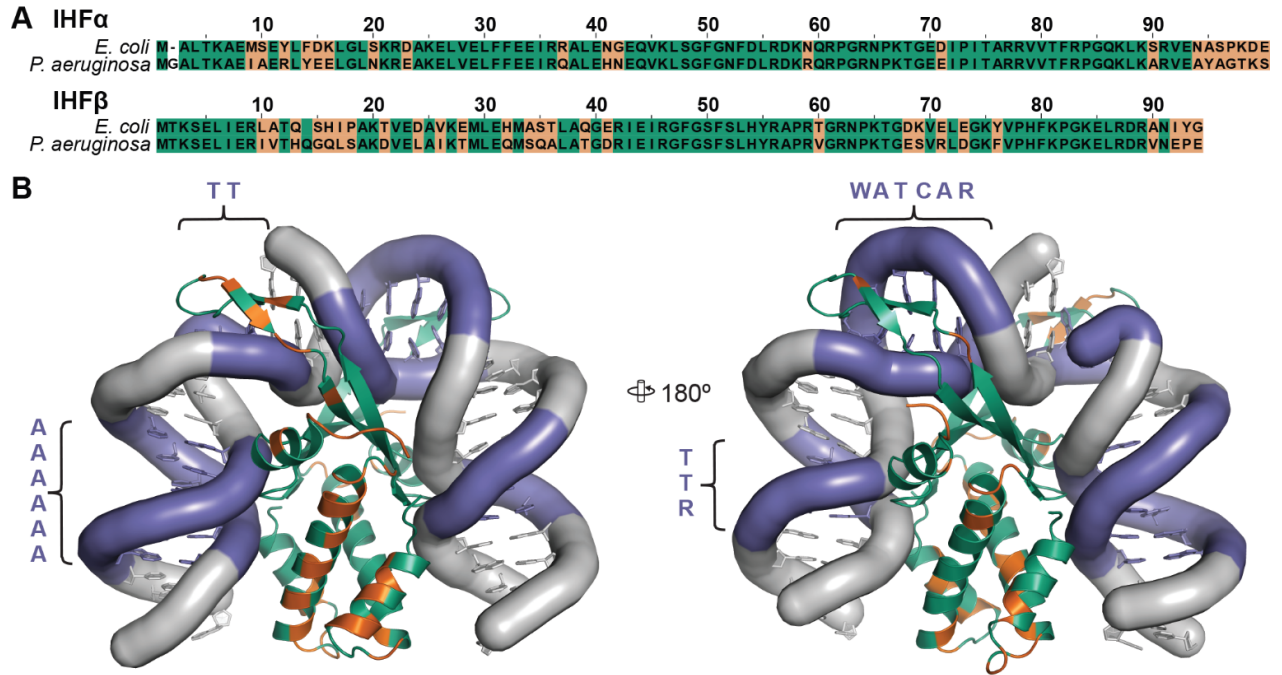


Figure S3.2. DNA-interacting residues are conserved between *E. coli* and *P. aeruginosa* IHF proteins. Related to Figure 3.3.

(A) Pairwise alignments of *E. coli* and *P. aeruginosa* IHF α (top), and IHF β (bottom). Identical residues (green) and non-identical residues (orange) are indicated. (B) Cartoon representations of *E. coli* IHF bound to cognate DNA (PDB:1IHF). DNA consensus sites (purple) are labelled. Non-consensus DNA regions are shown in light grey. Both IHF monomers are shown in green. Residues that are not identical in *P. aeruginosa* IHF monomers are shown (orange), none of which are expected to affect clamping of the A-tract, clamping of the TTR motif, kinking of DNA, or recognition of the TT and WATCAR motifs.

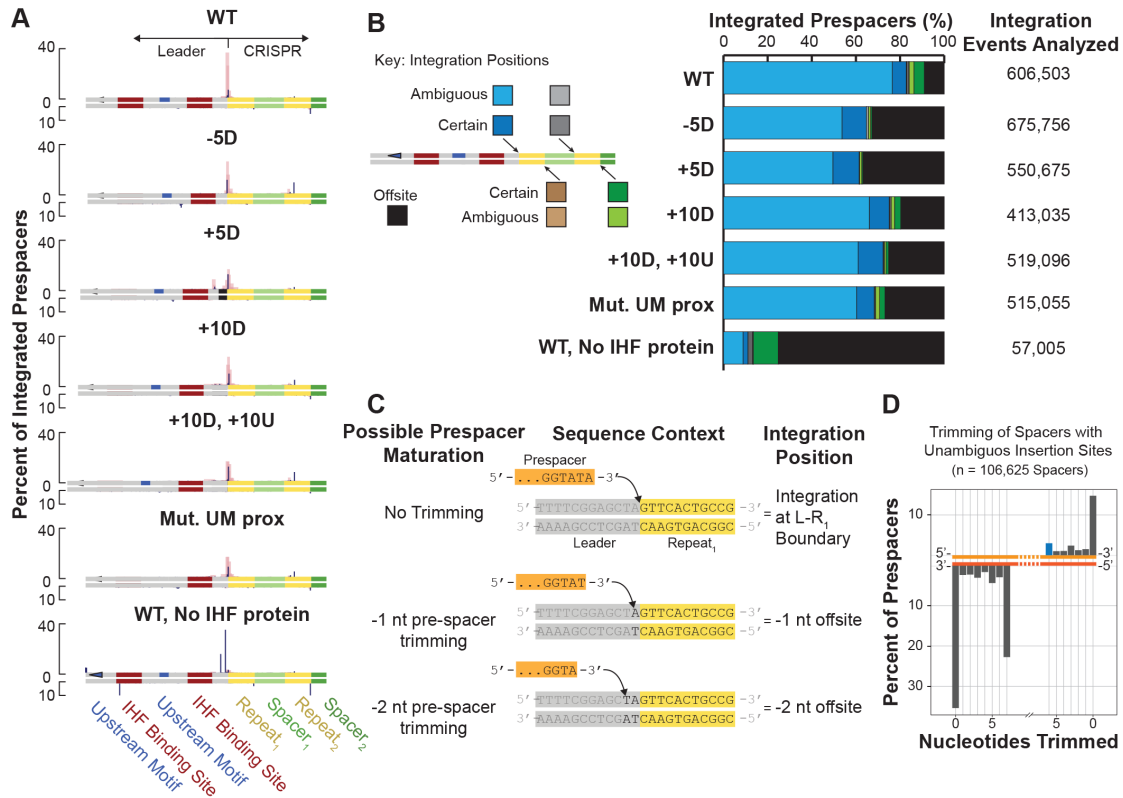


Figure S3.3. High-throughput sequencing and quantification of prespacer integration events. Related to Figures 3.3 and 3.4.

(A) Graphical representation of Cas1-2/3-mediated integration events of a 40 bp dsDNA prespacer into a wildtype fragment of the CRISPR2 locus from *P. aeruginosa* PA14, in the absence of IHF (WT, No IHF protein), or in the presence of IHF (WT). Integration into variants containing leader mutations was also analyzed. “-5D” signifies deletion of 5 bp downstream of IHF proximal site. “+5D” signifies the addition of 5 bp downstream of IHF proximal site. “+10D” signifies the addition of 10 bp downstream of IHF proximal site. “+10D,+10U” signifies the addition of 10 bp both downstream and upstream of the IHF proximal site. “Mut. UM prox” is used to describe a CRISPR containing a scrambled proximal upstream motif. Prespacers with confirmed integration positions (blue) represent the minority of mapped reads, since most integration positions cover two to three nucleotides that could have come from either the prespacer or the CRISPR substrate (red boxes). Based on this uncertainty, we refer to these insertions as “ambiguous”. (B) Data from panel A is quantified (right). Reads were counted as “ambiguous” insertion events at leader-repeat or spacer-repeat boundaries if the window of possible integration positions for that read covers one of the boundaries shown (left). (C) Ambiguous integration events arise because prespacers can be trimmed by Cas1-2/3 or contaminating nucleases prior to integration^{S2,S3}. An example of an ambiguous integration event at the leader-repeat₁ boundary is shown. (D) Quantification of trimming events in prespacers confirmed integration positions into WT CRISPR substrate. Half of the prespacers analyzed are trimmed (49.8%), but there is no preference for cleavage at a PAM-spacer boundary (position shown in blue) indicating that under the conditions tested, prespacer maturation may be stochastic.

Uncropped images of denaturing gel electrophoresis of integration reactions into a ^{32}P -labelled CRISPR fragment substrate. Reactions were performed in the absence of Cas1-2/3 (left), in the presence of all components (middle), or in the absence of IHF (right). Each row of gels depicts a replicate.

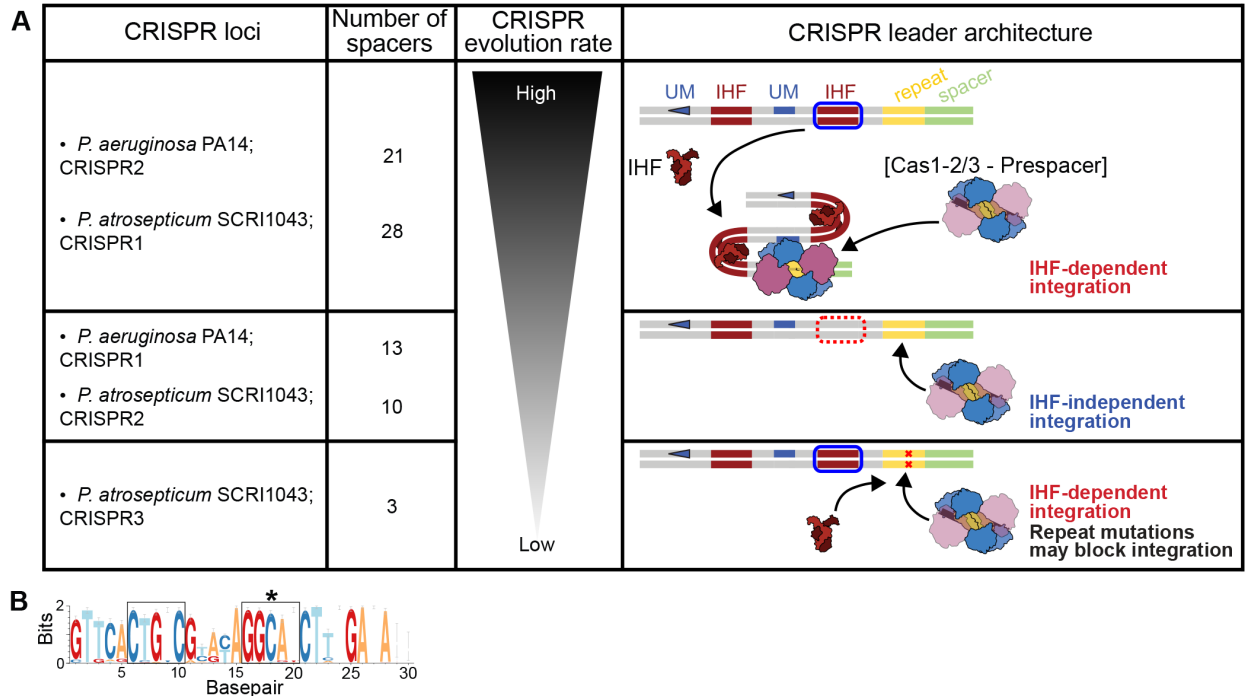


Figure S3.5. Summary of differential evolution rates between co-existing CRISPR loci. Related to Figure 3.3.

(A) CRISPR loci found in *P. aeruginosa* PA14 and *P. atrosepticum* SCRI1043 fall into three categories based on the number of spacers, and rates of evolution. These categories correspond with three different CRISPR architectures. Rapidly evolving CRISPRs from these bacteria possess leaders with a proximal IHF site (midpoint 28 bp from the L-R junction), and a conserved upstream motif (midpoint 63 bp from the L-R junction) (top). IHF folds the leaders of these loci, to recruit an upstream motif to a specific surface of Cas1-2/3, facilitating efficient integration. Slower evolving CRISPRs do not have recognizable IHF binding sites in their leaders but do retain a correctly positioned upstream motif (middle). Integration at these loci is less efficient, and proceeds via an IHF-independent pathway. The slowest evolving CRISPR between these model organisms has the correct motifs in its leader, but integration may be blocked by a rare mutation within the inverted repeat of the first CRISPR repeat (bottom). Mutations in these regions of the CRISPR repeat are known to strongly inhibit integration. (B) Inverted repeat within I-F CRISPR repeat is highly conserved. Sequence logo of the first repeat in I-F CRISPRs. The highly conserved inverted repeats, necessary for efficient spacer integration, are boxed. The position of the C18T mutation that occurs in the first repeat of the *P. atrosepticum* CRISPR3 locus is indicated with an asterisk.

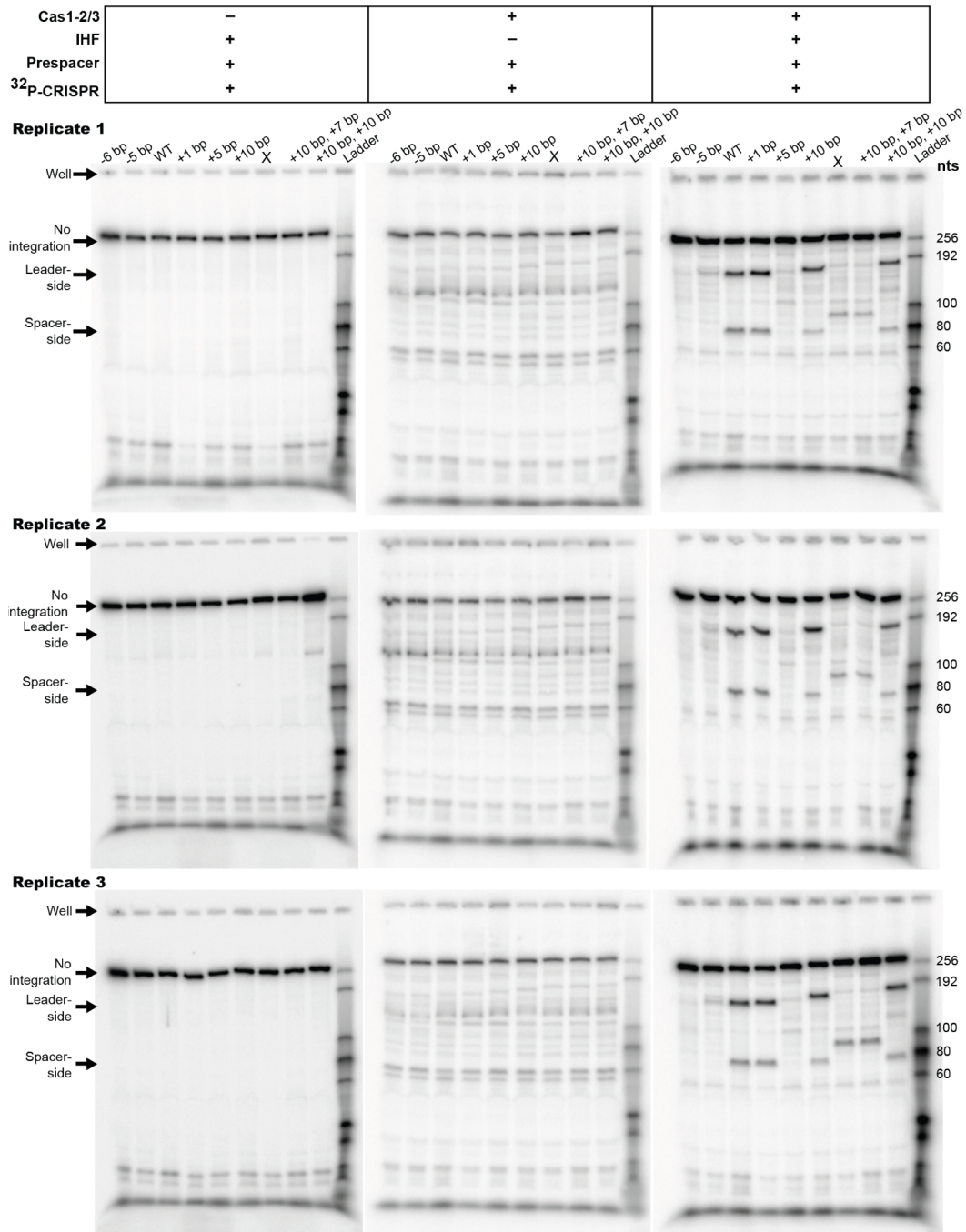


Figure S3.6. Leader motifs involved in CRISPR evolution are phase sensitive. Related to Figure 3.4.

Uncropped images of denaturing gel electrophoresis of integration reactions into a ³²P-labelled CRISPR fragment substrate. Reactions were performed in the absence of Cas1-2/3 (left), in the absence of IHF (middle), or in the presence of all components (right). “X” marks lanes containing reactions that were not examined further and are not discussed in this paper. Each row of gels depicts a replicate.

Materials and Methods

Bacterial Strains

Escherichia coli DH5 α (Thermo Fisher Scientific) cells were used to amplify plasmids used in this paper. *E. coli* BL21 DE3 (NEB) cells were used to express proteins used in this paper. *E. coli* were grown in LB (Lennox) media at either 37°C, or 16°C after induction of protein expression with 0.5 mM IPTG (isopropyl- β -D-thiogalactoside), shaking in baffled conical flasks at 200 rpm. The type I-F CRISPR system Cas proteins and CRISPR loci were cloned from *P. aeruginosa* UCBPP-PA14.

Modelling the I-F CRISPR Integration Intermediate Complex

I-F CRISPR leaders contain an 8 bp insertion between the leader-repeat junction and the proximal IHF site relative to IHF site-containing I-E CRISPR leaders. To model the impact of this insertion, 8 bp of idealized B-form dsDNA containing 11 bp/turn was built using the 3D-DART web server⁶¹. This 8 bp duplex was inserted immediately downstream of the IHF proximal binding site in the I-E Cas1-2-IHF-CRISPR holo-complex structure (PDB: 5WFE) using PyMOL v1.8.2.3 (Schrödinger, LLC). Next, components of the I-F Cas1-2/3 adaptation complex were docked onto the I-E Cas1-2 adaptation complex. Docking was performed by superimposing the I-F Cas1 dimers (PDB: 3GOD) onto each I-E Cas1 dimer (PDB: 5WFE) using the “cealign” command in PyMOL v1.8.2.3 (Schrödinger, LLC). The Cas1 homodimers superimpose with a root mean squared deviation of 1.3 Å over 144 α -carbon atoms. I-F Cas2/3 proteins (PDB: 5B7I) were similarly modeled by superimposing one copy of the I-F Cas2 domain of Cas2/3 onto each of the I-E Cas2 subunits. The Cas2 subunits superimpose with a root mean squared deviation of 1.2 Å over 36 α -carbon atoms. Images of the resulting models were rendered using ChimeraX v1.1^{62,63}.

Building a Database of CRISPR Leaders

A total of 15,567 complete bacterial genomes and 2,658 complete bacterial chromosomes were downloaded from the NCBI RefSeq Assembly database on June 10th of 2019. In addition, 351 complete archaeal genomes and 25 complete archaeal chromosomes were downloaded from the NCBI GenBank Assembly database on the same day. CRISPRDetect v2.4²⁰ was used to identify CRISPR loci in all downloaded genomes. Search parameters were set using the following command, “-word_length 11 -minimum_word_repeatation 3 -max_gap_between_crisprs 125 -repeat_length_cutoff 17 -minimum_repeat_length 23 -minimum_no_of_repeats 3 -check_direction 1 -array_quality_score_cutoff 3”. The first repeat and 200 nucleotides upstream of each CRISPR locus (leader) were collected for downstream analyses. Regions corresponding to both sides of the CRISPR were downloaded when CRISPRDetect was not able to reliably determine the directionality (i.e., distinguish the leader from the trailer). Leaders were assigned to a particular CRISPR subtype by CRISPRDetect v2.4, or by proximity to subtype-specific *cas* genes. The remaining 13% of CRISPR loci (1,996 loci) could not be annotated by CRISPRDetect and lacked nearby *cas* genes.

Identification of Conserved DNA Motifs in CRISPR Leaders

CRISPR leaders were analyzed using the MEME webserver with default settings⁶⁴. Custom position weight matrices for I-E leader IHF binding sites, I-F leader IHF binding sites, I-E leader upstream motifs, I-F leader upstream motifs, I-C upstream motifs, II-C upstream motifs, I-E_A, I-E_B, and the I-E leader-anchoring motif, and the FIMO-identified sequences found in CRISPR leaders have been submitted to the PRODORIC2 webserver and are available in <https://doi.org/10.1016/j.cub.2021.05.068>. A local copy of FIMO²² was used to find significant

matches to each of the position weight matrices. A minimum p-value threshold of $1E-4$ was used to define significant position weight matrices matches. The position weight matrix for the I-F leader IHF binding site does not represent the entire protein-occluded region of DNA. Therefore, tallies for the I-F IHF binding site were extended to represent the 29 bps of DNA that are occluded by bound IHF²⁸, in order to calculate IHF binding site midpoints. This analysis allows for comparison of IHF binding site distributions between CRISPR leaders, and for comparison in the position of motifs mapped to the sense and anti-sense strands. Sequence logos for matched DNA sequences were generated with a local copy of WebLogo 3.7⁶⁵.

The microbial genome dataset used here was downloaded from the NCBI RefSeq database on June 10th of 2019. As of April 1st of 2021, an additional 8,703 bacterial genomes, 1,277 bacterial chromosomes, 67 archaeal genomes and 6 archaeal chromosomes have been added to the NCBI RefSeq database. These additions represent about a third of the total number of currently available genomes and chromosomes, with large numbers of genomes originating from uncultured bacteria and the Phyla *Planctomycetes*, *Ignavibacteria* and *Candidatus Saccharibacteria*. The exclusion of these new genomes is not expected to change the conclusions of this paper, which focuses on CRISPR subtypes (I-E, I-F, I-C and II-C) predominantly found in *Actinobacteria*, *Proteobacteria* and *Firmicutes*⁶⁶. However, we expect that additional bioinformatic studies will uncover more leader motifs that regulate CRISPR adaptation or transcription.

Phylogenetic Analyses

CRISPR leaders containing sequences matching any of the position weight matrices described in this paper, were fetched using a Python v2.7 script that uses the Bio SeqIO package. A non-redundant list of CRISPR leaders was generated using CD-HIT v4.8.1^{67,68} with a 95%

identity cutoff. Non-redundant CRISPR leaders were then aligned with a local version of MAFFT v7.429⁶⁹, using the following command-line options: “--genafpair --maxiterate 1000 --thread 100 --threadit 100 --threadtb 100”. The resulting alignment was then analyzed with MaxAlign v1.1⁷⁰ to find and remove misaligned or non-homologous sequences which introduced a large number of gaps in the alignment. The list of remaining leader sequences were then realigned as above. A maximum-likelihood phylogenetic tree was then generated from realigned leader sequences, using FastTree 2.1.11⁷¹. The tree was generated using the following parameters “-gamma -spr 4 -mlacc 2 -slow -nt”. Trees were visualized in RStudio by overlaying the data of motif distances using ggtree⁷²⁻⁷⁴ ggplot2⁷⁵ and tidyverse⁷⁶.

Plasmid Construction

The *cas1* and *cas2/3* genes from *P. aeruginosa* UCBPP-PA14 have been previously cloned into a spectinomycin-resistant p2S LIC vector²⁵ (Addgene, #89240). The *ihfA* and *ihfB* genes from *P. aeruginosa* UCBPP-PA14 were PCR-amplified to construct an N-terminal 6x-Histidine tagged variant of IHF α , an untagged IHF β and an N-terminal Strep-II tagged variant of IHF β . HRV3C protease cleavage sites were included between the affinity tags and the protein sequence of interest. 6x-His-IHF α and untagged IHF β were cloned into MCS1 and MCS2 of pACYCDuet-1, respectively. This plasmid (pHisIHF α IHF β) has been deposited with Addgene (#149384). StrepII-IHF β was cloned into pET28a (pStrepIHF β) and has been deposited with Addgene (#149385). The *P. aeruginosa* UCBPP-PA14 CRISPR2 locus (position 2935917-2937205, in the genome NC_008463), as well upstream leader DNA (311 bp) and downstream terminus DNA (303 bp) were PCR-amplified and cloned into pHERD30T⁷⁷, (pCRISPR2_wt, Addgene #149386). Leader sequence variants were generated via site directed mutagenesis using pCRISPR2 as template,

followed by ligation of the linearized plasmids. All variants are deposited at Addgene, and are listed in the Key Resources Table (pCRISPR2_-6D, #149387; pCRISPR2_-5D, #149388; pCRISPR2_+1D, #149389, pCRISPR2_+5D, #149390; pCRISPR2_+10D, #149391; pCRISPR_+10D+7U, #149393; pCRISPR_+10D+10U, #149394; pCRISPR2_IHF_Opt, #149395; pCRISPR2_IHF_Mut, #149396; pCRISPR2_motif_scram, #149397; pCRISPR2_IHFdist_rm, #162318; pCRISPR2_IRdist_scram, #162319).

IHF Expression and Purification

E. coli BL21(DE3) cells were co-transformed with both pHisIHF α -IHF β and pStrepIHF β . Cells were grown in LB-Miller media (10 g/L Tryptone, 10 g/L NaCl, 5 g/L yeast extract), supplemented with 34 μ g/mL Chloramphenicol and 50 μ g/mL Kanamycin, at 37°C and 200 rpm to an OD₆₀₀ of 0.45. Cultures were then cooled on ice for one hour, without agitation, before the addition of 0.2 mM IPTG. Cells were then grown for an additional 18 hours at 16°C, before centrifugation at 5000 g for 10 minutes. Cell pellets were suspended in Lysis Buffer (25 mM HEPES-NaOH pH 7.5, 500 mM NaCl, 10 mM Imidazole, 1mM TCEP, 5% Glycerol) supplemented with 0.3x HaltTM Protease Inhibitor Cocktail (ThermoFisher), and sonication at 30% amplitude for a total on-time of 6 minutes (1 second on-time with 3 seconds off), at 4°C. Lysate was clarified by two successive centrifugations at 12,000 rpm for 15 minutes each at 4°C. Clarified lysate was then flowed over HisTrap HP resin (Cytiva), to affinity purify His-tagged IHF heterodimers. HisTrap HP resin was washed with 20 column volumes of IHF Lysis Buffer containing 50 mM Imidazole. His-tagged IHF heterodimers were then eluted from the HisTrap HP resin with Lysis Buffer containing 500 mM Imidazole. IHF was then concentrated at 4°C (Corning Spin-X concentrators). 6x-Histidine and StrepII tags were cleaved from IHF α and IHF β monomers

respectively, in the presence of PreScission protease, overnight at 4°C while dialyzing into Lysis Buffer containing no Imidazole. Remaining 6x-His-IHF α and PreScission protease were removed by affinity chromatography using HisTrap HP resin (Cytiva). The IHF heterodimer was concentrated and diluted in buffer to reduce the NaCl concentration to 200 mM. Untagged IHF heterodimer was then purified by affinity chromatography on Heparin Sepharose (GE Healthcare), which was washed with 10 column volumes of Heparin Wash Buffer (25 mM HEPES-NaOH pH 7.5, 200 mM NaCl, 5% Glycerol). IHF was eluted from HiTrap Heparin HP resin (Cytiva) with a linear gradient to Heparin Wash Buffer containing 2 M NaCl. Fractions containing IHF heterodimer were concentrated at 4°C (Corning Spin-X concentrators), before purification on a Superdex 75 size-exclusion column (Cytiva) in Heparin Wash Buffer.

Cas1-2/3 Expression and Purification

E. coli BL21(DE3) cells were transformed with pCas1Cas23 and grown 2x 1 L LB-Miller media (10 g/L Tryptone, 10 g/L NaCl, 5 g/L yeast extract), supplemented with 50 μ g/mL Spectinomycin, at 37°C and 200 rpm to an OD₆₀₀ of 0.45. Cultures were then cooled on ice for two hours, without agitation, before being induced with 0.2 mM IPTG. Cells were then grown for an additional 18 hours at 16°C, before being centrifuged at 5000 g for 10 minutes. Each cell pellet, originating from 1 L of culture, was resuspended in 20 mLs of Cas1-2/3 Lysis Buffer (50 mM HEPES pH 7.5, 500 mM KCl, 10% Glycerol) supplemented with 0.3x Halt™ Protease Inhibitor Cocktail (ThermoFisher). Cells were lysed via sonication, and lysate was clarified as above. StrepII-tagged Cas1-Cas2/3 was affinity purified on StrepTrap HP resin (GE Healthcare) and eluted with Cas1-2/3 Lysis Buffer containing 3 mM desthiobiotin (Sigma-Aldrich). Eluate was concentrated at 4°C (Corning Spin-X concentrators), before purification over a Superdex 200 size-

exclusion column (Cytiva) equilibrated in 10 mM HEPES pH 7.5, 500 mM KCl, and 10% Glycerol.

Nucleic Acid Preparation

To generate a splayed prespacer with a double-stranded core, sense (5'-[Alexa546]TACATGCTCTAGCAAAACGACTTGACACAACGAGG) and antisense (5'-AAATTAAGTGCAAGTCGTTTTGCTAGAGCTACAT) DNA strands (Europhins) were resuspended in Hybridization Buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM MgCl₂) and mixed in equimolar amounts. The bases corresponding to the double-stranded core are underlined. These ssDNAs were annealed to each other by denaturation at 95°C for 5 minutes, followed by slowly cooling to room temperature over 1 hour. Annealed splayed prespacer was purified by electrophoresis through a 8% (w/v) (29:1 mono:bis) polyacrylamide in 1x TBE (100 mM Tris-Borate pH 8.3, 2 mM EDTA) at 4°C. A band corresponding to splayed prespacer was excised from the gel and purified by ethanol precipitation. A 40 bp dsDNA prespacer was made in a similar manner from sense (5'-TCTACATGGTCTAGGAAAAGGACTTGGACAAGGAGGTATA-3') and antisense (5'-TATACCTCCTTGTCCTTTTCCTAGACCATGTAGA-3') strands (Europhins).

To make ³²P-labelled CRISPR integration substrates, primers (Forward primer: CCAATTGCCCCGAAGCTTC-3'; Reverse primer: 5'-TCCAGAAGTCACCACCCG-3') (Europhins) complementary to far upstream in the leader and to within the 2nd spacer of the CRISPR loci were used to amplify a DNA fragment containing most of the leader and the beginning of the CRISPR locus, from pCRISPR plasmid variants (deposited with Addgene) as templates. These PCR products were purified on a 2% (w/v) agarose native gel and extracted with

a gel DNA recovery kit (Zymo Research). 1 pmole of dsDNA, corresponding to 2 pmoles of 5' ends, was end-labelled on both strands with 4 pmoles of [γ - 32 P]ATP (PerkinElmer) by polynucleotide kinase (NEB) in 1x PNK buffer at 37°C for 45 minutes. PNK was heat-denatured by incubation at 65°C for 20 minutes. Spin column purification (G-25, GE Healthcare) was used to remove unincorporated radioactive nucleotides and to buffer exchange DNAs into 1x TE (10 mM Tris-HCl pH 8, 1 mM EDTA).

In vitro Integration Assays

Integration reactions were performed using 300 nM of splayed prespacer (purified as described above), 200 nM Cas1-2/3, ~1 nM of 32 P-labelled CRISPR variant fragment and 350 nM IHF, in Integration Buffer (20 mM HEPES pH 7.5, 150 mM Potassium acetate, 5 mM MnCl₂, 1 mM DTT, 5% Glycerol), for 1 hour at 37°C. Reactions were stopped by the addition of SDS to 1% and subsequent phenol-chloroform extraction. The nucleic acid containing layer was mixed 1:1 with 2x formamide loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) and denatured at 95°C for 5 minutes, before resolving full-length 32 P-labelled CRISPR strands from those fragmented by an integration event on a 8% (w/v) (29:1 mono:bis) polyacrylamide Urea gel in 1x TBE. Gels were dried and quantified using a Typhoon phosphorimager (GE Healthcare). The intensity of full-length CRISPR variant, leader-side integration fragments, spacer-side integration fragments, non-specific integration events, and background readings were quantified with Multi Gauge v3 (Fujifilm). We then calculated the no integration, leader-side integration, spacer-side integration and non-specific integration events, as percentages of total events.

High-throughput Sequencing of *In vitro* Integration Products

Integration reactions were performed using 200 nM of a 40 bp dsDNA prespacer (purified as above), 200 nM Cas1-2/3, 1 nM of CRISPR variant fragment, and 350 nM IHF, in Reaction Buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 5 mM MnCl₂, 1 mM DTT, 5% Glycerol), for 1 hour at 37°C. Reactions were then phenol-chloroform extracted, and then further purified with a DNA Clean and Concentrator Kit (Zymo Research). 1 µL of eluted DNA was used as a PCR template. Combinations of four primers (P1, 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACACCCGGCTTTCTTAG-3'; P2, 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAATTGCCGAAGCTTC-3'; P3, 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGTCTAGGAAAAGGACTTGGAC-3'; P4, 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCCTTGTCCAAGTCCTTTTCC-3') were used to amplify all possible integration products and to simultaneously add Illumina adaptor sequences, using Q5 DNA polymerase (NEB). These PCR products were purified on a 1% (w/v) agarose gel and extracted with a gel DNA recovery kit (Zymo Research). DNA barcoding and paired-end sequencing were performed at the University of Montana Genomics Core, on an Illumina MiSeq 300 V2. Paired end reads were merged in PEAR⁷⁸ and aligned to integrated and CRISPR substrates in BLAST⁷⁹. Downstream sequence analysis was performed in RStudio. The region flanking an integration site often contained 1-3 nucleotides of homology to the end of a prespacer substrate, making it impossible to distinguish if the matching nucleotides came from the prespacer molecule or the CRISPR locus-containing substrate. In the latter event, the prespacer may have been trimmed prior to integration as was seen for about half of the spacers with a clearly discernible integration site. These integration events were therefore marked as 'ambiguous', with the range of true integration points spanning up to a three-nucleotide window (**Figure S3.3**).

Quantification and Statistical Analysis

A threshold p -value of 1E-4 was used to report significant matches of motif position weight matrices to leader sequences by FIMO²². Autocorrelation analysis of leader motifs was performed in OriginPro (OriginLab). Quantification of gel bands for integration experiments was performed in Multi Gauge (FUJIFILM) image analysis software, from 3 independent reactions and denaturing PAGE gel images. The mean $\pm$ 1 standard deviation is reported. In Figure 3.3, errors were propagated to report the fold integration measured in CRISPR variants relative to wildtype. In Figure 3.4 the raw mean was plotted $\pm$ 1 standard deviation and the data was fit to a sine wave of equation: $y = y_0 + A \times \sin(\pi \frac{x-x_c}{w})$ (in which y_0 is the y-axis offset, A is the amplitude, w is the period, and x_c is the phase shift). The R^2 of the fits were 0.95 for the leader-side integration data and 0.75 for the spacer-side integration data.

CHAPTER FOUR

IDENTIFICATION OF SUBTYPE- AND TAXON-SPECIFIC CRISPR LEADER MOTIFS ENHANCES CRISPR DETECTION AND ANNOTATION

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Murat Buyukyoruk

Contributions: Designed the research, performed the research, prepared manuscript figures, wrote the manuscript.

Co-Author: Pushya Krishna

Contributions: Designed the research, performed the research.

Co-Author: Andrew Santiago-Frangos

Contributions: Designed the research, wrote the manuscript.

Co-Author: Blake Wiedenheft

Contributions: Designed the research, wrote the manuscript.

Manuscript Information

Murat Buyukyoruk, Pushya Krishna, Andrew Santiago-Frangos, Blake Wiedenheft

Status of Manuscript:

X Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

IDENTIFICATION OF SUBTYPE- AND TAXON-SPECIFIC CRISPR LEADER MOTIFS
ENHANCES CRISPR DETECTION AND ANNOTATION

Murat Buyukyoruk¹, Pushya Krishna¹, Andrew Santiago-Frangos¹, Blake Wiedenheft¹

¹Department of Microbiology and Cell Biology, Montana State University, Bozeman MT
59717, USA.

Abstract

Bacteria and archaea acquire resistance to genetic parasites by preferentially integrating short fragments of foreign DNA at one end of a CRISPR. “Leader” DNA upstream of the CRISPR regulates the integration of new foreign DNAs into the first CRISPR repeat, and the transcription of the CRISPR. We analyze the leaders of 24,940 CRISPRs from 25,900 microbial genomes to identify conserved sequence motifs. A global analysis of all leader sequences fails to identify universally conserved motifs. However, an analysis of leader sequences that have been grouped by 16S rRNA-based taxonomy and CRISPR subtype reveals 77 specific motifs in type I, II, III and V CRISPR leaders. Several leader motifs coincide with sequences previously reported to regulate CRISPR integration, transcription, and processing. Most identified leader motifs are novel with unknown function, although their conservation suggests important biological roles. We show that leader motifs improve detection and classification of CRISPR loci.

Introduction

The repeat-spacer-repeat DNA sequence of CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) first noted in 1987, stores DNA-based memories of past phage infections as unique “spacer” sequences (**Figure 4.1a**)¹⁻⁴. CRISPR adaptive immunity begins with Cas (Crispr-associated) proteins integrating new foreign DNA fragments at the first repeat of a CRISPR, as a new spacer sequence^{5, 6}. The CRISPR locus is then transcribed into a single RNA strand that is processed by Cas proteins into individual crRNAs (Crispr RNAs)⁷. Sequence motifs in the DNA upstream of the CRISPR (leader) regulate integration at the first CRISPR repeat, transcription of the CRISPR locus, and processing of the first crRNA⁸⁻¹⁵. Finally, the crRNAs guide Cas surveillance complexes (i.e., Cas9) to bind and degrade complementary DNA or RNA sequences, destroying the foreign genetic element⁷.

Large scale bioinformatic analyses have identified thousands of CRISPR-Cas systems in bacterial and archaeal genomes, which can be broadly classified into six types, and over 33 subtypes¹⁶. Each CRISPR system type is characterized by unique biochemical activities, highlighting the evolutionary arms race between phages and bacteria. Over the past decade the wide array of biochemical activities of CRISPR-Cas systems has been repurposed for revolutionary applications in gene editing, diagnostics, and DNA and RNA recording devices^{17, 18}. Therefore, we anticipate that new CRISPR-Cas discovery methods will provide new insights into the natural diversity of bacterial and archaeal adaptive immune systems and will lead to novel biotech applications. Current CRISPR-Cas discovery methods rely on either the identification and annotation of the repetitive CRISPR array sequence, or of subtype-specific *cas* genes^{16, 19-22}. However, repeat-based CRISPR annotation struggles to annotate short or degenerate CRISPRs,

such as the atypical CRISPRs found in Tn7-CAST (CRISPR-associated transposon) systems^{23, 24}. Whereas Cas-based annotations have no utility for annotating the roughly 25% of bacterial and archaeal CRISPRs that are found far away from *cas* genes²⁵.

Most CRISPR adaptive immune systems consist of three types of DNA sequences: i) *cas* genes, ii) CRISPR array, iii) leader (**Figure 4.1a**), and yet the leader has not been previously utilized for CRISPR-Cas annotation^{16, 19-22}. We previously analyzed 15,274 CRISPRs and identified conserved leader motifs that correlated with taxonomy and CRISPR subtype¹³. Type I-C, I-E, I-F and II-C leaders that occurred in *Pseudomonadota* had IHF binding sites and subtype-specific motifs, whereas type I-E leaders in *Actinomycetota* had novel motifs¹³. Based on these observations, we hypothesized that conserved CRISPR leader motifs could be used to annotate CRISPR-Cas systems. To test this hypothesis, we analyzed leader sequences from 24,940 high-confidence CRISPRs originating from 25,502 bacterial and 398 archaeal genomes. This analysis reveals 77 leader motifs that are specific to microbial taxonomy and CRISPR subtype, and which occur at discrete distances from the leader-repeat junction. Some of these motifs coincide with sequences previously reported to regulate CRISPR integration, transcription, and processing, however most are novel conserved motifs that likely have critical roles in CRISPR biology¹⁰⁻¹⁵. We show that leader-based annotations perform as well as repeat-based annotations for high-confidence CRISPRs, and leader-based annotations outperform repeat- and Cas-based annotations for short or degenerate CRISPRs. Overall, these data show that leader diversity mirrors CRISPR and Cas diversity, leader sequences co-evolve with CRISPR-Cas systems and host factors, and leader-based annotations improve CRISPR discovery.

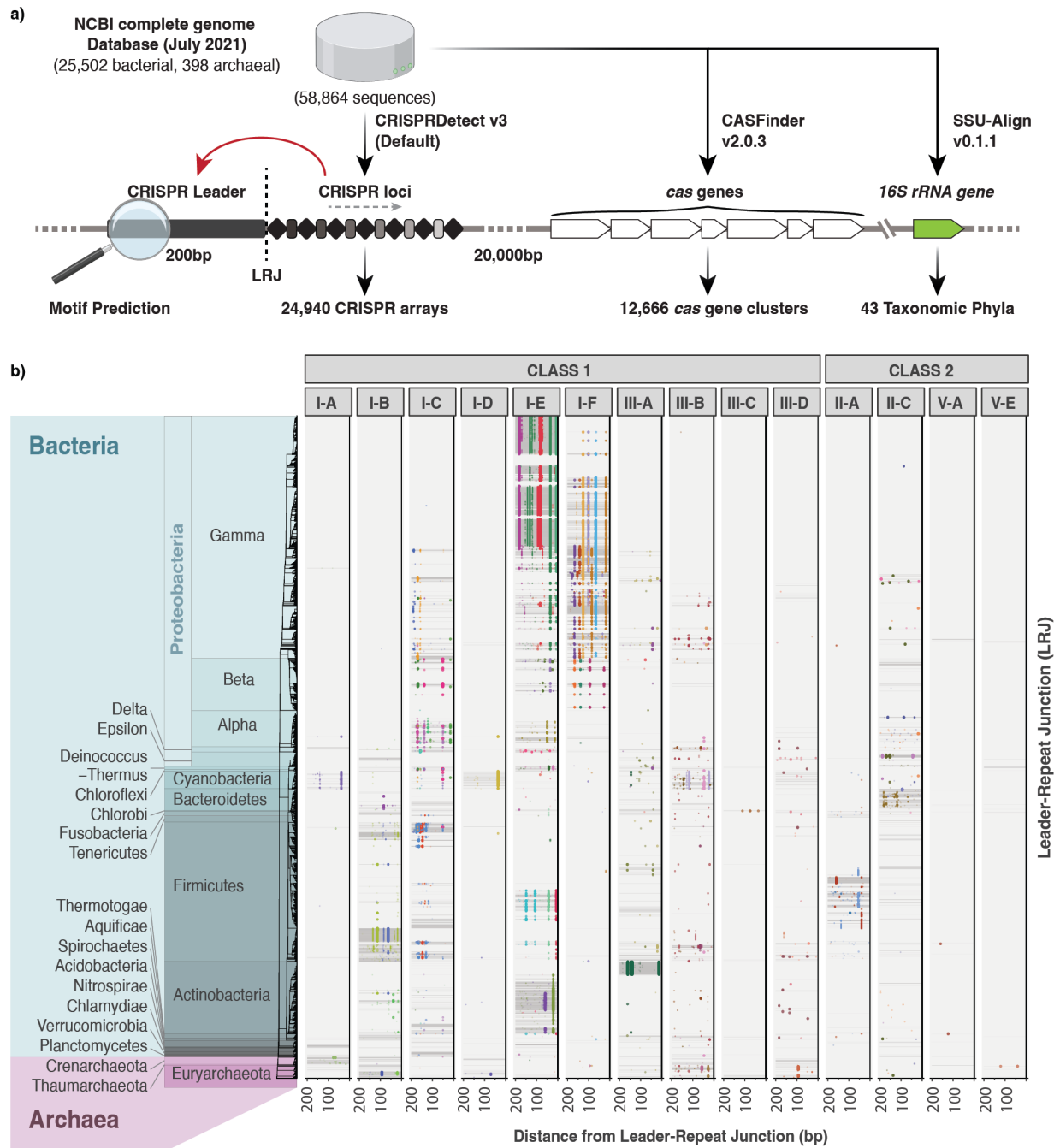


Figure 4.1. Subtype and taxonomy-specific leader motifs are widespread in bacteria and archaea. **(a)** Schematic representation of CRISPR-Cas locus and bioinformatic software used to identify genetic features. Typical CRISPR-Cas locus in bacteria and archaea consist of three types of DNA sequences including *cas* genes, CRISPR array and CRISPR leader. **(b)** Distribution of 77 motifs predicted in CRISPR leaders. Maximum-likelihood phylogenetic tree of 16S rRNA sequences identified from 25,502 bacterial and 398 archaeal genomes. Genomes with multiple CRISPR

arrays are represented with duplication of terminal branches according to the number of CRISPR arrays found in genomes. A total of 41,007 branches are shown. Shaded gray lines in each subtype panel represent the presence of a subtype-specific CRISPR array in the corresponding genome that classified according to repeat and *cas*-based approaches. Each dot represents the midpoint of motifs identified from subtype/taxonomy paired leader sequences. Different colored dots represent different motifs. The size of each dot is adjusted according to the calculated final motif score (see methods). Many of the type I leaders (except I-U) are shown to possess leader motifs found between 0 and 200bp upstream of the leader-repeat junction (LRJ). Subtype- and taxon-specific are shown to be found in many type III leaders (III-A, III-B, III-C, and III-D) and some type II (II-A and II-C) and type V (V-A and V-E) leaders.

Results

Identification and Taxonomic Distribution of 77 CRISPR Leader Motifs

Previous studies have suggested that leader sequences co-evolve with subtype-specific Cas proteins and phyla-specific host factors to facilitate integration^{12, 14, 26}. Therefore, we hypothesized that CRISPR leader sequences grouped by CRISPR subtype and host taxonomy would be enriched in specific motifs that would assist CRISPR annotation. To test this hypothesis, we first used existing repeat- and Cas-based annotations to generate a list of 19,767 high-confident CRISPRs from 38 bacterial and 7 archaeal phyla and 25,900 genomes (**Figure 4.1a and Table S4.1**). Although leaders can be up to 500 bp in length, the majority of CRISPR leaders appear to be shorter than 200 bps²⁷. To limit introducing noise for most leaders, and to apply the same methodology to all sequences, we analyzed the first 200 bps upstream of 19,767 CRISPRs. The leader sequences were then grouped by CRISPR subtype and 16S rRNA-based taxonomy of the host genome, resulting in an initial list of 264 leader groups that contained at least one sequence, out of 1,075 possible subtype-taxonomy combinations. Next, we removed highly similar (>60% identity) leader sequences within each group, to assist the identification of conserved motifs. This curation step had the greatest effect on the I-E Gammaproteobacteria leaders, decreasing the

number of sequences in this group from 6,843 to 220, which was likely due to an overrepresentation of *Escherichia coli* genomes in the NCBI database. We then used MEME to identify 124 motifs in 107 groups of leaders that consisted of a minimum of six sequences²⁸. We had previously noticed that leader motifs involved in integration appeared at strongly conserved positions within the leader¹³. Therefore, the list of 124 candidate motifs was inspected manually, and a final list of 77 motifs that occur in discrete locations upstream of the CRISPR was generated (**Figure S4.1**). The number of motifs identified per CRISPR type correlated with the abundance of leaders per CRISPR type, such that 45 motifs were identified from type I leaders, 19 motifs from type III, 11 motifs from type II, 2 motifs from type V, and 0 motifs from type IV and VI (**Figure S4.2**).

To determine whether these motifs are restricted to subtype-taxa groups of leader sequences, we used FIMO to identify statistically significant matches to these motifs in the leaders of all 24,940 original CRISPRs (**Figure 4.1b**)²⁹. This analysis reveals that the final list of 77 motifs is predominantly specific to subtype-taxa groups of sequences. Further, these 77 motifs occur in discrete locations upstream of the CRISPR, either invariant in their location (seen as columns of colored dots), or as shifting by a multiple of 10-11 bp (preserving the phase of the motif along the helical structure of dsDNA) (**Figure 4.1b and Figure S4.1**).

Leader Motifs Complement Repeat- and Cas-based Annotations of CRISPRs

Repeat- and Cas-based annotations identify the subtype of 19,767 out of the 24,940 CRISPRs (79.24%) that were picked by CRISPRDetect. We hypothesized that leader motifs could be used to enhance the number and accuracy of CRISPR array subtype annotations. To test this hypothesis, we developed a scoring matrix that accounted for the presence of multiple motifs

expected to co-occur, the expected base pair position of a given motif based on the CRISPR leader training dataset, and the statistical significance of the motif match (**Figure 4.2a**).

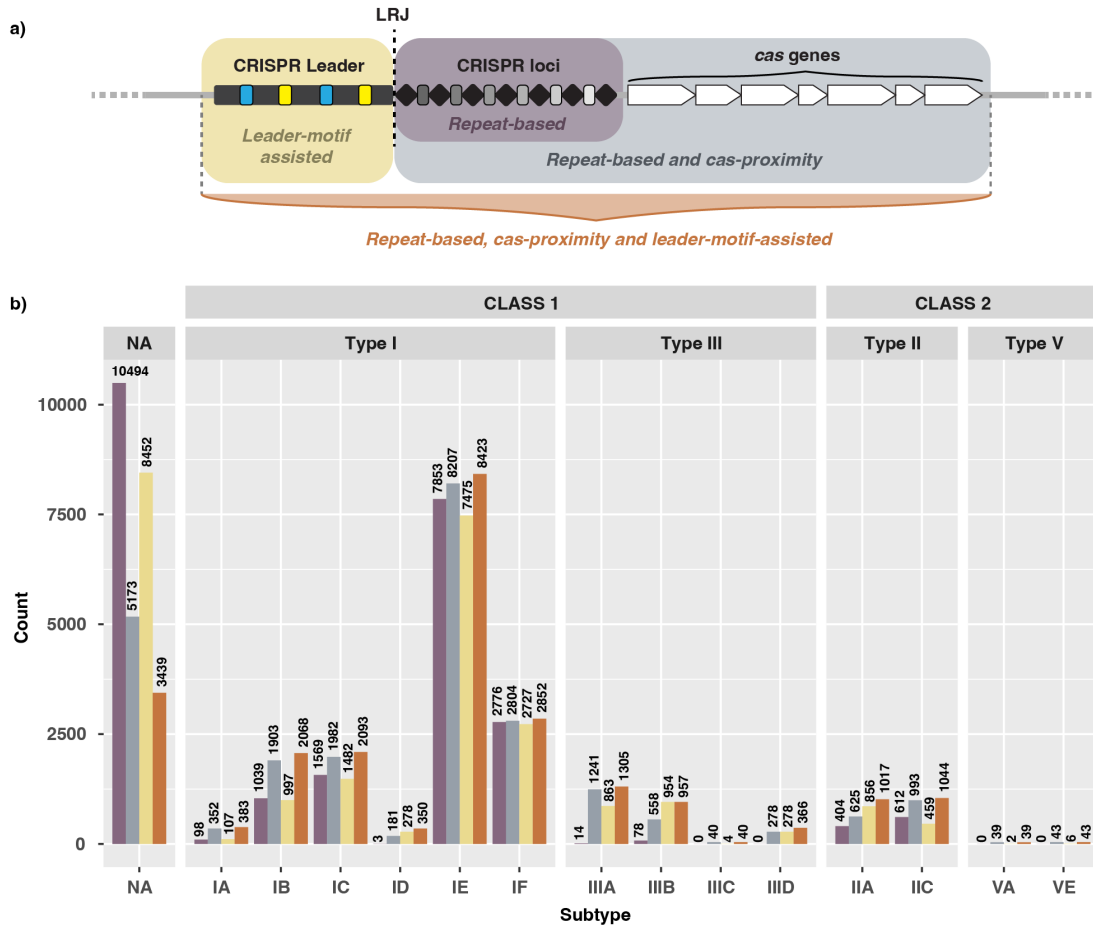


Figure 4.2. Leader motifs complement repeat- and Cas-based annotation of CRISPRs.

(a) Scheme of repeat-, Cas-, and leader-based methods for annotating CRISPRs. **(b)** Performance of CRISPR annotation methods. The number of high confidence loci through leader-based annotation (score > 0.05) is increased in type I, II and III, while type V has more limitations with a few new array predictions. Leader-based CRISPR annotation coupled with traditional annotation methods classified 67.2% (7,055) of CRISPRs not annotated by repeat-based methods, and classified 33.5% (1,734) of CRISPRs not annotated by a combination of repeat- and cas-based methods.

Leader-based annotations outperformed repeat-based annotations for most CRISPR subtypes, and complemented repeat-based annotations to increase the number of classified CRISPRs by 2,042 (from 14,446 to 16,488). Further, leader motifs complemented the combined repeat- and Cas-based annotations to increase the number of classified CRISPRs by 1,734 (from 19,767 to 21,501) (**Figure 4.2b**).

To examine the accuracy of this CRISPR classification strategy, we measured the correlation of leader-based annotation to the original repeat- and Cas-based annotations of leaders. Leader-based annotations agreed with 82.74% and disagreed with 1.84% of repeat-based annotations. Leader motifs could not be used to annotate the remaining 15.42% of CRISPRs because of insufficient example CRISPRs in rare subtype-taxa groups (**Figure S4.3b**). An expected drawback was that leader-based annotations could not identify type III-C, V-A and V-E CRISPRs, due to an insufficient number of these loci in the leader motif identification training set. To our surprise, leader-based annotations show lower correlation (69.98%) with CRISPRs that had been annotated by both repeat similarity and proximity to *cas* genes (<20 Kb from a subtype-specific gene). This perceived misclassification by leader motifs may be real, or may be due to multiple Cas surveillance complexes sharing the same CRISPR locus, as is common for some co-occurring type I and III or IV systems³⁰⁻³⁴ (**Figure S4.3a**). As an additional control to measure the false-positive rate of leader-based annotations, we generated three individual test sets of 1,000 randomly generated 200-bp long sequences. Less than 3% of all randomly generated sequences had a region with a significant match to a leader motif, and less than 0.6% of all randomly generated sequences had a significant match to a leader motif at the expected base pair position, resulting in a total of 27 final false annotations from 3,000 random sequences (**Figure S4.4**). Collectively these data

show that CRISPR leader sequences contain information that can be used for accurate annotation of CRISPR arrays.

CRISPR Leader Motifs Enable Annotation of Short and Degenerate Putative CRISPR Arrays

A textbook example of a CRISPR is composed of many identical repeat sequences. However, many biologically relevant CRISPR loci are composed of few repeats, or repeats that are imperfect copies of each other, or both^{23, 24}. We hypothesized that leader-based annotations could enhance our ability to distinguish short and degenerate CRISPR arrays from non-CRISPR repetitive elements. To test this hypothesis, we first modified CRISPRDetect to export putative CRISPR arrays that are typically scored as “questionable” and not shown to the user. In addition to 24,940 high-confidence CRISPRs, CRISPRDetect identified 82,116 putative CRISPRs. Of the 82,116 putative CRISPRs, the repeat- and Cas-based annotations classify 1,911 CRISPRs, while leader-based annotations classify an additional 8,301 CRISPRs (10,212 total).

We then examined which new CRISPRs are identified by this approach. Most of the newly identified CRISPR arrays are I-E, I-F, I-B and I-C subtypes, while only a few are type III-C, V-A and V-E subtypes, which coincides with the bias in the original CRISPR dataset used to train the leader-based annotation method (**Figure 4.3**). Most of the newly classified CRISPRs (8,523 of 10,212) are short and consist of two or three repeats, and 5,987 of these CRISPRs are unique (**Figure 4.4a**). The leader-based annotation approach identifies CRISPRs that appear to have short repeats and long spacers (**Figure 4.4b,c**). However, closer inspection of these sequences reveals that the repeat sequences are degenerate, which causes CRISPRDetect to report a short consensus repeat sequence, and append the excluded repeat nucleotides onto the spacer sequences (**Figure S4.5b,c**).

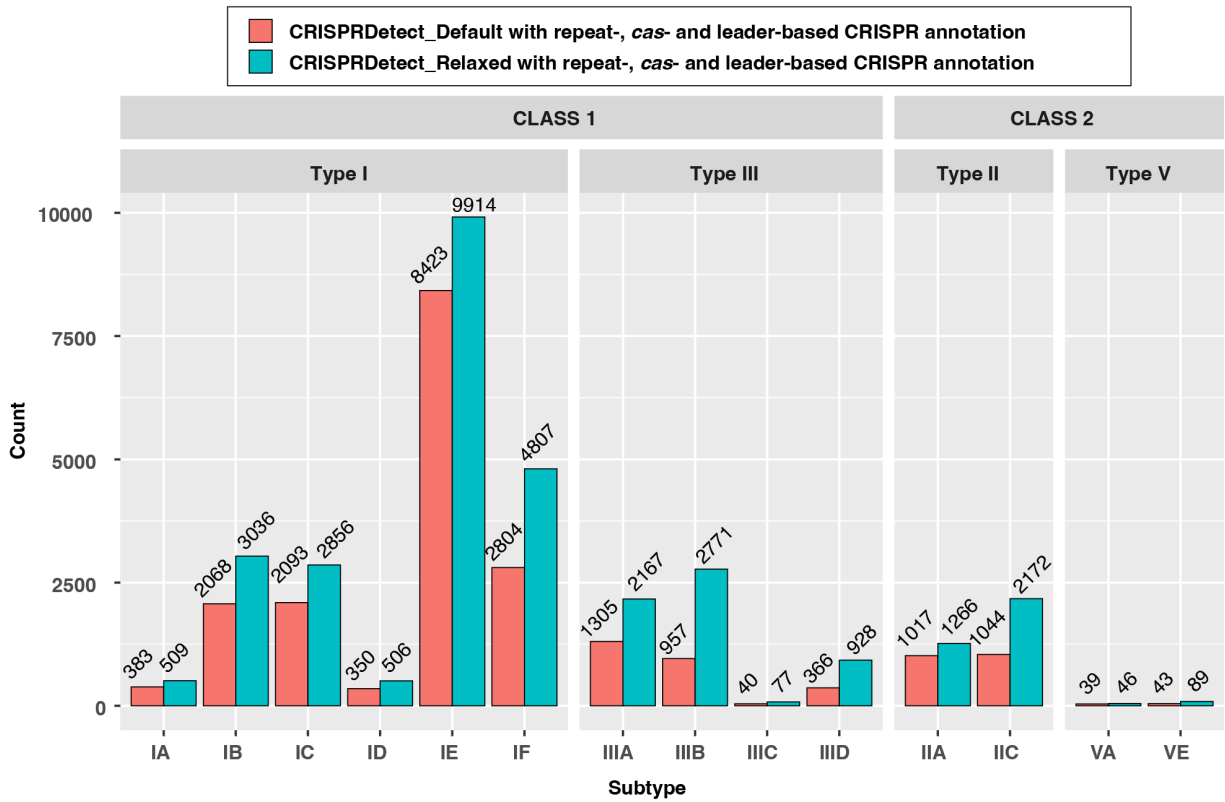


Figure 4.3. Leader-based annotation improves the classification of short and degenerate CRISPR arrays.

Leader-based annotation of 107,056 putative arrays predicted by an edited version of CRISPRDetect to identify short and degenerated arrays. Bar plot shows an increased detection of CRISPR arrays across several subtypes. The number of annotated arrays were increased from 20,760 to 31,144 with the edited version of CRISPRDetect coupled with leader motif-assisted annotation.

Leader-based annotations were trained on CRISPRs from bacterial and archaeal genomes, but we wondered whether these motifs could be used to classify viral CRISPR sequences. The ICP1 phage encodes a compact type I-F CRISPR system³⁵. CRISPRDetect identifies 79 putative CRISPRs (24 unique leader and CRISPR sequences, and 5 unique leader sequences) in 69 ICP1 phage genomes. 24 of these ICP1 phage CRISPR arrays are unique with different spacers, suggesting they are actively acquiring new spacer sequences. Repeat-based annotations classify 26/79 of these CRISPRs as type I-F subtype. However, leader-based annotations correctly classify

71/79 of ICP1 phage CRISPRs as belonging to the type I-F subtype and the *Pseudomonadota* taxa (Table S4.3)³⁵. These annotations are consistent with the ICP1 phage possessing a type I-F Cas integrase and surveillance complex, and to its host bacterium *Vibrio cholera* belonging to the phylum *Pseudomonadota*.

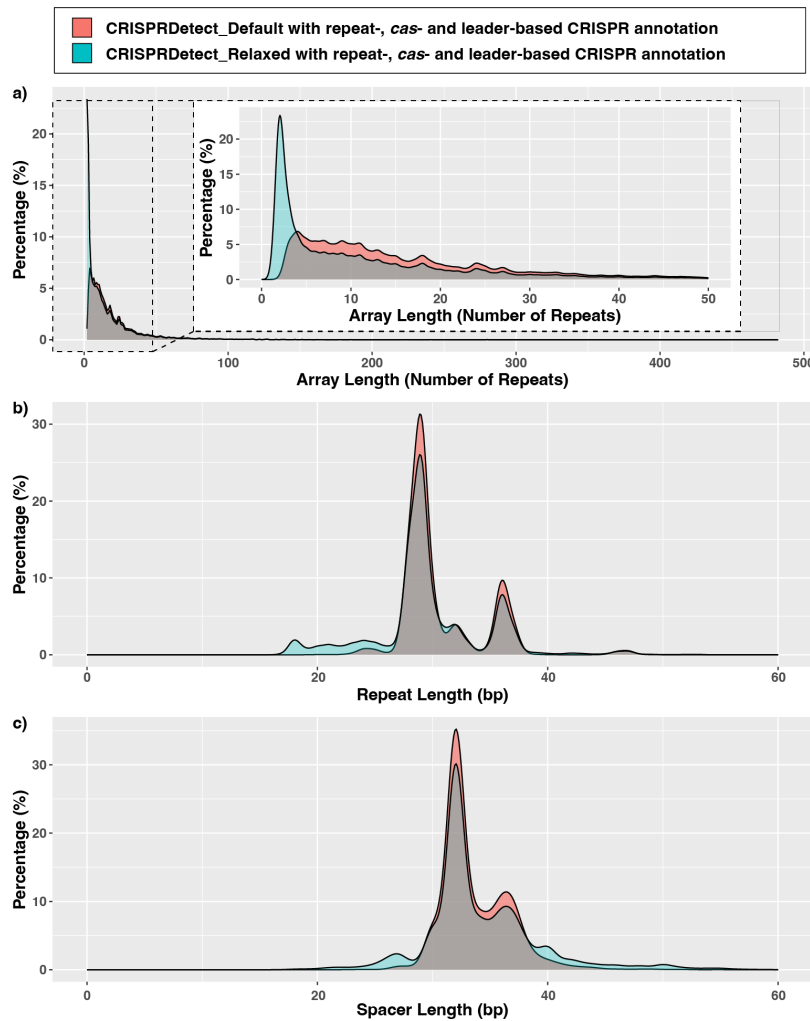


Figure 4.4. Distribution of array, repeat and spacer lengths predicted by an edited version of CRISPRDetect coupled with leader-based annotation.

Arrays predicted by default CRISPRDetect parameters (orange) and edited CRISPRDetect parameters (teal) are shown. **(a)** Plot of the length of classified CRISPR arrays. CRISPRs shorter than 50 repeats are shown in the subpanel. **(b)** Plot of the length of repeats in classified CRISPRs. **(c)** Plot of the length of spacers in classified CRISPRs.

Collectively these analysis shows that leader-based annotations enable the annotation of CRISPR arrays that have few or imperfect repeat sequences, and can be used to annotate bacterial, archaeal and viral CRISPRs.

Discussion

Here we demonstrate that diverse CRISPRs have conserved leader sequence motifs that can be repurposed for accurate annotation of CRISPR loci. Many leader motifs are subtype-specific, and therefore leader-based annotations effectively complement traditional repeat- and Cas-based annotations to classify short and degenerate CRISPRs, orphan CRISPRs that lack nearby *cas* genes, or CRISPRs that occur near multiple *cas* genes from different subtypes. Conversely, several leader motifs are taxa-specific because of host factor signatures, and therefore may be used to determine recent horizontal gene transfer events of CRISPR-containing DNAs, determine the taxonomy of CRISPR-containing genome fragments, or identify the host organism of CRISPR-containing viruses^{11, 13, 36}. In a head-to-head comparison, leader-based annotation coupled to CRISPRDetect identification of putative CRISPR loci, outperformed other CRISPR annotation methods by uniquely identifying 6,452 CRISPRs (**Figure S4.6**).

Leader motif discovery was most successful for CRISPR subtypes which were well-represented in the training set of leaders used to identify motifs, namely I-A, I-B, I-C, I-D, I-E, I-F, III-A, III-B, III-D, II-A and II-C. Our training set is lacking many examples of leaders from roughly 20 CRISPR subtypes, therefore we correspondingly did not identify leader motifs for most CRISPR subtypes. We anticipate that many more discoveries and improvements to leader-based annotations await as repeat- and cas-based annotations of CRISPRs improve. Further, to keep the methodology and analysis consistent, 200 bp of DNA was considered for every leader, although

there is evidence suggesting that archaeal leaders can range from 100-500 bp long^{27, 37}. Nevertheless, the analysis described here provides the most detailed and broad report of taxa- and subtype-specific leaders that we are aware of, reporting a total of 77 conserved motifs from 14 CRISPR subtypes across 22 bacterial and 3 archaeal taxa.

Beyond their utility in CRISPR classification, conserved leader sequence motifs play critical roles in CRISPR biology. An examination of the 77 leader motifs that we describe reveals several matches to DNA sequences that have been previously determined to be important for CRISPR integration, transcription, or processing (**Table S4.4**)¹⁰⁻¹⁵. Concerning integration, previous studies have shown that DNA binding sites for the taxa-specific host factor IHF, and for the subtype-specific CRISPR integrase complex, are required for integration at the first repeat of I-E and I-F CRISPRs¹¹⁻¹⁴. These integration-critical leader motifs are represented in part (i.e., *IE Gammaproteobacteria motif 4* which contains part of an IHF binding site), or in combination (i.e., *IE Gammaproteobacteria motif 2* which contains an “Upstream Motif” and part of an IHF binding site) (**Table S4.4**). A recent report has revealed that Archaeal histones are required for efficient and preferential integration of new spacers at the first CRISPR repeat in *Pyrococcus furiosus*³⁸. A single minimal 30 bp histone dimer binding site is represented in the motif “*IIIB Euryarchaeota motif 2*”, which matches to half of the most proximal histone tetramer binding site (**Table S4.4**)³⁸. As for transcription, the archaeal proteins Transcription Factor B and TATA binding protein bind to BRE (B recognition element) and TATA-box DNA sequences to initiate transcription³⁹. Previous studies have shown that leaders of I-A, I-B, I-D, III-B and III-D CRISPRs in *Euryarchaeota* and *Thermoproteota* have BRE and TATA-box motifs that are necessary for CRISPR transcription^{31, 32, 40}. The leader motif analysis captures some of these transcription-critical leader motifs in “*IIIB*

Euryarchaeota motif 1” which matches to the BRE and TATA-box, and “*IIID Euryarchaeota motif 1*” which matches to the TATA-box (**Table S4.4**). Finally, recent studies have shown that the transcribed region of the type II-A leader sequence in *Firmicutes* hybridizes with the first repeat and interacts with the second repeat, to promote processing of the first crRNA¹⁵. A portion of the leader critical for processing of the most recently acquired spacer is represented in “*IIA Firmicutes motif 2*” (**Table S4.4**). These examples highlight that identified motifs can coincide with a fragment of a larger biologically relevant sequence, or can represent multiple sequences. In sum, these comparisons show that the leader analysis workflow described here identifies biologically relevant DNA sequence motifs.

Leader- and repeat-based annotations provide an accurate and informative classification of CRISPR integration, transcription, and processing. However, there Cas-based annotations may still be required in some cases to provide a holistic view of the biological role of a given CRISPR locus. For example, several Type IV-A CRISPRs have repeat and leader sequences with high sequence identity to co-occurring type I-E CRISPRs, and have recently been shown to co-opt the type I-E Cas1-2 integration complex for new spacer acquisition^{33, 34}. Further, RNA-targeting type III-B Cas systems often repurpose the spacers from DNA-targeting type I-A CRISPRs to provide another layer of immune defense^{30, 31}. The re-use of spacers from the same CRISPR locus by multiple different Cas surveillance complexes may be a common phenomenon to launch a multipronged attack against a phage and bypass anti-CRISPRs.

Of the 77 conserved leader motifs that we identify, only nine have known biological roles (**Table S4.4**), highlighting a knowledge gap in our understanding of CRISPR, bacterial and archaeal biology. Therefore, the newly identified leader motifs reported here provide a springboard

for new studies that further increase our understanding of CRISPR biology, in the case of subtype-specific motifs, or bacterial and archaeal biology, in the case of taxa-specific motifs. Finally, we show that leader-based annotations enhance the identification and annotation of CRISPRs from bacterial, archaea and viruses.

Methods

Building a Database of CRISPR Leaders

A total of 25,502 complete bacterial and 398 archaeal genomes and chromosomes were downloaded from the NCBI RefSeq Assembly database (accessed on June 10th, 2021). CRISPRDetect v3²¹ with default parameters was used to identify CRISPR loci within 58,864 genomic and plasmid sequences, resulting in 24,940 high-confidence CRISPR loci predictions (array quality score >3). Two hundred nucleotides upstream of each CRISPR locus was extracted and both sides of the CRISPR locus were considered in cases where CRISPRDetect was unable to reliably determine directionality (n=374).

Subtype Annotation of CRISPR Arrays in CRISPR Leader Database

To classify 24,940 leaders into various CRISPR subtypes, CRISPRDetect was employed, as well as a further analysis to evaluate their proximity to subtype-specific *cas* genes. Subtype classification of CRISPR arrays by CRISPRDetect is determined when the repeats of an array match to repeats associated with a particular subtype. Using this approach, CRISPRDetect was used to annotate 57.9% (n=14,446) of the total CRISPRs. To assign subtype to the remaining 10,494 unclassified CRISPR arrays we used PRODIGAL v2.6.3⁴¹ to translate 20,000 nucleotides upstream and downstream of each unclassified CRISPR locus in all six frames. These open reading

frames were then queried for *cas* genes using MacsyFinder v1.0.5⁴² with the following parameters: “*macsyfinder --sequence-db <open_reading_frames> --db-type gembase -d <CRISPR_subtype_definitions> -p <HMM_profiles> -w 50 -vv all*”. HMM profiles and classification definitions (*CRISPR_subtype_definitions*) used in MacsyFinder are acquired from the local version of CRISPRCasFinder (v4.2.20, CasFinder v2.0.3)²⁰. Classification through the proximity to subtype-specific *cas* genes resulted in the annotation of 5,321 unclassified arrays, while the remaining 5,173 CRISPR arrays remained unclassified due to the lack of nearby *cas* genes.

Identification of Host-taxonomy

Taxonomizr v0.5.3⁴³ was used to identify the taxonomy of downloaded genomes. Taxonomy information was used to assign kingdom and phylum to each genome. 2,551 CRISPRs were found on plasmids. Taxonomic information for plasmids was assigned by linking the plasmid accession number to the corresponding host genome accession number.

Identification of Conserved DNA Motifs in CRISPR Leaders

CRISPR leaders were divided into subsets according to CRISPR subtype (type I-IV and subtypes A-G) and 16S rRNA-based taxonomy of the host (n=263 subsets). From each subset, a non-redundant list of 200 nt-long CRISPR leader DNA sequences was generated using UCLUST v11.0.667⁴⁴ with a 60% similarity cutoff. This cutoff was empirically determined through an iterative process designed to optimize motif identification. Motif discovery was performed for each subset using a local copy of MEME⁴⁵. The following parameters were used: “*meme -mod anr -nmotifs 6 -minw 6 -maxw 40 -allw -objfun se -maxiter 1000 -shuf 2 -nostatus -bfile background_file -hsfrac cutoff_value*”. The *background_file* is generated via the first-order Markov

model with the following command: *fasta-get-markov -dna -nostatus -nosummary -m 1*". The *cutoff_value* for *hsfrac* was set to 0.75, unless it was determined otherwise by MEME due to the low number of sequences available in the queried subset.

The finite amount of sequence data and the many degrees of freedom in a motif probability matrix can lead to the prediction of motifs with deceptively low and significant E-values. Therefore, we generated several examples of randomly shuffled sequences for each subtype- and taxon-specific group of CRISPR, which were used as input data for motif prediction, to establish a significant E-value cutoff for motifs predicted for unshuffled subtype- and taxon-specific groups of sequences. This process was repeated ten times, with the average of the lowest E-values from each run being adopted as a threshold. Consequently, Position Weight Matrices (PWMs) anticipated from each subset exceeding this predetermined cutoff were dismissed, resulting in 124 significant motifs. Next we manually inspected distribution of motifs within CRISPR subtype- and taxon-specific groups of leaders, and removed motifs that occurred in a broad number of positions upstream of the leader-repeat junction. The final 77 motifs, featuring narrow distributions in CRISPR leader sequences, were selected to compile a list of subtype and taxa-specific leader motifs.

Scoring Matrix for Leader-motif-assisted CRISPR Annotation

We incorporate 77 CRISPR leader motifs into previously established CRISPR prediction software to enhance the classification of CRISPR arrays missed by traditional detection methods. To minimize false positive detection of motifs in CRISPR leader sequences, a scoring system was developed that took into account the i) significance (p-value) of the matched sequence to the PWM of the motif, ii) frequency that a given motif was observed to occur in leaders during the motif

identification workflow, and iii) the number of subtype-specific motifs observed in a given CRISPR leader. A local copy of FIMO v5.0.5²⁹ was used to identify significant matches of each PWM in all CRISPR leaders (observed). To generate a reference dataset, an additional FIMO search was also performed using the non-redundant CRISPR leader subsets where the motif was initially identified (expected). The p-value of motif occurrence in a CRISPR leader ($p_{observed}$) was compared against the lowest p-value ($p_{expected}$) from the reference dataset. Next, a position score was calculated according to the distance between the leader-repeat-junction (LRJ) and the midpoint of the matching sequence (x). The number of motifs observed at the calculated position ($d_{observed, 0 < x < 200}$) in the reference dataset was divided by the number of motifs observed with the most abundant distribution ($d_{max, 0 < x < 2000}$). The distribution fraction is multiplied by the fraction calculated according to the p-value to determine a motif-specific score (Equation 1).

$$Score_{Motif} = score_{pval} * score_{dist} = \frac{\log(p_{observed})}{\log(p_{expected})} * \frac{\sum(d_{observed, 0 < x < 200})}{\sum(d_{max, 0 < x < 2000})}$$

Equation 1

The calculated score was assigned to each motif occurrence in CRISPR leaders. In cases where a single CRISPR leader contains multiple motifs representing the same subtype, the final motif score was summed and used to assign a classification to the CRISPR arrays (Equation 2). Leaders with a final score lower than 0.05 remained unclassified to prevent misclassification of arrays.

$$\sum_{i=1}^{n_{subtype-motif}} Score_{Motif,i}$$

Equation 2

16S Ribosomal RNA Analysis of Bacterial and Archaeal Genomes with CRISPR

SSU-align v0.1.1⁴⁶ was used to predict 16S rRNA sequences from 25,502 complete bacterial genomes and 398 archaeal genomes. 16S rRNA sequences were then aligned with a local version of MAFFT v7.429⁴⁷ using the following command-line options: “*--globalpair --maxiterate 1000*”. 16S rRNA sequences were duplicated according to the number of CRISPR arrays predicted within the organisms (i.e., each CRISPR loci is associated with a 16S sequence). CRISPR arrays found in plasmids were also linked to the host 16S rRNA sequence. A maximum-likelihood phylogenetic tree was then generated from the 41,007 16S rRNA sequences using FastTree v2.1.10⁴⁸. The tree was generated using the following parameters “*-gamma -spr 4 -mlacc 2 -slow -nt*”. Trees were visualized in RStudio by overlaying the data of motif distances using ggtree^{49, 50} and ggplot2⁵¹. Midpoints of the predicted motifs are plotted as dots. Each dot representing the motif is sized according to the calculated motif score.

Identification of Short and Degenerate Arrays

An edited version of CRISPRDetect is used to capture putative CRISPR arrays that contain short and degenerate CRISPRs. The source code of CRISPRDetect was edited to export putative CRISPR arrays that are typically excluded because these arrays scored poorly, have short repeat sequences, or have long spacer sequences. The following CRISPRDetect parameters were used with the edited version of CRISPRDetect: “*-word_length 11 -minimum_word_repeatation 2 -max_gap_between_crisprs 125 -repeat_length_cutoff 17 -minimum_repeat_length 17 -minimum_no_of_repeats 2 -check_direction 1 -left_flank_length 500 -right_flank_length 500 -check_plasmid 1*”. A total of 107,056 putative arrays predicted from the edited version of CRISPRDetect. These arrays were then subjected to traditional and leader-motif-assisted

annotation, resulting in the classification of 31,144 arrays into CRISPR subtypes. Compared with the hits obtained by the default version of CRISPRDetect, 10,212 arrays of the predicted 31,144 arrays were previously eliminated by the default version or not identified as CRISPRs because they contain fewer than three repeats in the putative array.

Acknowledgements

We thank members of the Wiedenheft laboratory for comments, suggestions, and critical feedback. Research in the Wiedenheft laboratory is supported by National Institutes of Health (P20GM103500, P30GM110732-03, R01GM110270, and R01GM108888), National Science Foundation (EPSCoR EPS-110134), the M. J. Murdock Charitable Trust, and the Montana Agricultural Experimental Station. A.S-F. is a postdoctoral fellow of the Life Science Research Foundation that is supported by the Simons Foundation. A.S-F. is supported by the Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund. This work was supported by National Institutes of Health, United States grant 1K99GM147842 (A.S-F.).

References

1. Ishino, Y., M. Krupovic, and P. Forterre, *History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology*. J Bacteriol, 2018. **200**(7).
2. Mojica, F.J., et al., *Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria*. Mol Microbiol, 2000. **36**(1): p. 244-6.
3. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes*. Molecular Microbiology, 2002. **43**(6): p. 1565-1575.
4. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes*. Science, 2007. **315**(5819): p. 1709-12.
5. Pourcel, C., G. Salvignol, and G. Vergnaud, *CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies*. Microbiology (Reading), 2005. **151**(Pt 3): p. 653-663.
6. Lee, H. and D.G. Sashital, *Creating memories: molecular mechanisms of CRISPR adaptation*. Trends Biochem Sci, 2022. **47**(6): p. 464-476.
7. Nussenzweig, P.M. and L.A. Marraffini, *Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria*. Annu Rev Genet, 2020. **54**: p. 93-120.
8. Pul, U., et al., *Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS*. Mol Microbiol, 2010. **75**(6): p. 1495-512.
9. Pougach, K., et al., *Transcription, processing and function of CRISPR cassettes in Escherichia coli*. Mol Microbiol, 2010. **77**(6): p. 1367-79.
10. Phok, K., et al., *Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon Pyrococcus abyssi*. BMC Genomics, 2011. **12**: p. 312.
11. Nunez, J.K., et al., *CRISPR Immunological Memory Requires a Host Factor for Specificity*. Mol Cell, 2016. **62**(6): p. 824-833.
12. Wright, A.V., et al., *Structures of the CRISPR genome integration complex*. Science, 2017. **357**(6356): p. 1113-1118.
13. Santiago-Frangos, A., et al., *Distribution and phasing of sequence motifs that facilitate CRISPR adaptation*. Curr Biol, 2021. **31**(16): p. 3515-3524 e6.
14. Santiago-Frangos, A., et al., *Protein-mediated genome folding allosterically enhances site-specific integration of foreign DNA into CRISPRs*. 2023, bioRxiv.

15. Liao, C., et al., *Spacer prioritization in CRISPR-Cas9 immunity is enabled by the leader RNA*. Nat Microbiol, 2022. **7**(4): p. 530-541.
16. Makarova, K.S., et al., *Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants*. Nat Rev Microbiol, 2020. **18**(2): p. 67-83.
17. Santiago-Frangos, A., et al., *CRISPR-Cas, Argonaute proteins and the emerging landscape of amplification-free diagnostics*. Methods, 2022. **205**: p. 1-10.
18. Wang, J.Y., P. Pausch, and J.A. Doudna, *Structural biology of CRISPR-Cas immunity and genome editing enzymes*. Nat Rev Microbiol, 2022. **20**(11): p. 641-656.
19. Nethery, M.A., et al., *CRISPRclassify: Repeat-Based Classification of CRISPR Loci*. CRISPR J, 2021. **4**(4): p. 558-574.
20. Couvin, D., et al., *CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins*. Nucleic Acids Res, 2018. **46**(W1): p. W246-W251.
21. Biswas, A., et al., *CRISPRDetect: A flexible algorithm to define CRISPR arrays*. BMC Genomics, 2016. **17**: p. 356.
22. Crawley, A.B., J.R. Henriksen, and R. Barrangou, *CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems*. CRISPR J, 2018. **1**(2): p. 171-181.
23. Petassi, M.T., S.C. Hsieh, and J.E. Peters, *Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons*. Cell, 2020. **183**(7): p. 1757-1771 e18.
24. Klompe, S.E., et al., *Evolutionary and mechanistic diversity of Type I-F CRISPR-associated transposons*. Mol Cell, 2022. **82**(3): p. 616-628 e5.
25. Shmakov, S.A., et al., *CRISPR Arrays Away from cas Genes*. CRISPR J, 2020. **3**(6): p. 535-549.
26. Xiao, Y., et al., *How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration*. Nature, 2017. **550**(7674): p. 137-141.
27. Alkhnbashi, O.S., et al., *Characterizing leader sequences of CRISPR loci*. Bioinformatics, 2016. **32**(17): p. i576-i585.
28. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching*. Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
29. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif*. Bioinformatics, 2011. **27**(7): p. 1017-8.

30. Yair, Y. and U. Gophna, *Repeat modularity as a beneficial property of multiple CRISPR-Cas systems*. RNA Biol, 2019. **16**(4): p. 585-587.
31. Norais, C., et al., *Diversity of CRISPR systems in the euryarchaeal Pyrococcales*. RNA Biol, 2013. **10**(5): p. 659-70.
32. Nickel, L., et al., *Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III*. RNA Biol, 2013. **10**(5): p. 779-91.
33. Pinilla-Redondo, R., et al., *Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids*. Nucleic Acids Research, 2019. **48**(4): p. 2000-2012.
34. Benz, F., et al., *Type IV-A3 CRISPR-Cas systems drive inter-plasmid conflicts by acquiring spacers *in trans**. bioRxiv, 2023: p. 2023.06.23.546257.
35. Seed, K.D., et al., *A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity*. Nature, 2013. **494**(7438): p. 489-91.
36. Medvedeva, S., et al., *Virus-borne mini-CRISPR arrays are involved in interviral conflicts*. Nat Commun, 2019. **10**(1): p. 5204.
37. Rollie, C., et al., *Prespacer processing and specific integration in a Type I-A CRISPR system*. Nucleic Acids Res, 2018. **46**(3): p. 1007-1020.
38. Watts, E.A., et al., *Histones direct site-specific CRISPR spacer acquisition in model archaeon*. Nat Microbiol, 2023. **8**(9): p. 1682-1694.
39. Wenck, B.R. and T.J. Santangelo, *Archaeal transcription*. Transcription, 2020. **11**(5): p. 199-210.
40. Lillestol, R.K., et al., *CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties*. Mol Microbiol, 2009. **72**(1): p. 259-72.
41. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification*. BMC Bioinformatics, 2010. **11**: p. 119.
42. Abby, S.S., et al., *MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems*. PLoS One, 2014. **9**(10): p. e110726.
43. Sherrill-Mix, S., *taxonomizr: Functions to Work with NCBI Accessions and Taxonomy*. 2023.
44. Prasad, D.V.V., S. Madhusudanan, and S. Jaganathan. *uCLUST - A NEW ALGORITHM FOR CLUSTERING UNSTRUCTURED DATA*. 2015.

45. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
46. Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments*. Bioinformatics, 2009. **25**(10): p. 1335-7.
47. Katoh, K. and D.M. Standley, *MAFFT: iterative refinement and additional methods*. Methods Mol Biol, 2014. **1079**: p. 131-46.
48. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix*. Mol Biol Evol, 2009. **26**(7): p. 1641-50.
49. Yu, G., et al., *Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree*. Mol Biol Evol, 2018. **35**(12): p. 3041-3043.
50. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures*. Curr Protoc Bioinformatics, 2020. **69**(1): p. e96.
51. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.

Supplemental Tables and Figures

Class	Type	Subtype	CRISPR array Count	
			CRISPRDetect Annotation	CRISPRDetect Annotation w/ <i>cas</i> -gene proximity
Class 1	Type I	I-A	98	352
		I-B	1,039	1,903
		I-C	1,569	1,982
		I-D	3	181
		I-E	7,853	8,207
		I-F	2,776	2,804
		I-U	ND	274
	Type III	III-A	14	1,241
		III-B	78	558
		III-C	ND	40
III-D		ND	278	
Type IV	IV-A	ND	8	
Class 2	Type II	II-A	404	625
		II-B	ND	41
		II-C	612	993
	Type V	V-A	ND	39
		V-B	ND	6
		V-E	ND	43
		V-F	ND	14
		V-J	ND	59
		V-K	ND	3
	Type VI	VI-A	ND	10
		VI-B1	ND	93
		VI-B2	ND	2
		VI-C	ND	7
Unclassified		N/A	10,494	5,177
Total			24,940	24,940

Table S4.1. Subtype distribution of microbial CRISPRs.

24,904 CRISPR loci were identified in bacterial and archaeal genomes. CRISPRDetect annotation depends on array annotation according to matching repeat sequence the known CRISPR family repeats. Guilt-by-association analysis performed by searching *cas* genes with previously established *cas*-HMMs in six-frame translated 20kb upstream and downstream sequences of identified arrays to assign annotations to unclassified CRISPR arrays. (ND: Not Detected)

	Number of Motifs Found																								
Phylum	IA	IB	IC	ID	IE	IF	IU	IIIA	IIIB	IIIC	IIID	IVA	IIA	IIB	IIC	VA	VB	VE	VF	VJ	VK	VIA	VIB1	VIB2	VIC
Acidithiobacillia																									
Acidobacteria																									
Actinobacteria		1			2			1																	
Alphaproteobacteria			3		1										2										
Aquificae																									
Bacteroidetes		1													2										
Balneolaeota																									
Betaproteobacteria			2		2	3		1																	
Caldiserica																									
Calditrichaeota																									
CandidatusCloacimonetes																									
CandidatusKorarchaeota																									
CandidatusThermoplasmatota																									
Chlamydiae																									
Chlorobi										1															
Chloroflexi																									
Chrysiogenetes																									
Coprothermobacterota																									
Crenarchaeota	1																								
Cyanobacteria	1			1					3																
Deferribacteres																									
Deinococcus-Thermus					2																				
Deltaproteobacteria			1		1				1																
Dictyoglomi																									
Elusimicrobia																									
Epsilonproteobacteria								1			1				2										
Euryarchaeota		2		1					2		1							1							
Fibrobacteres																									
Firmicutes		2	2		3				1		2			3		1									
Fusobacteria																									
Gammaproteobacteria			2		4	6		2	2						2										
Ignavibacteriae																									
Kiritimatellaeota																									
Nitrospirae																									
Oligoflexia																									
Planctomycetes																									
Spirochaetes																									
Synergistetes																									
Tenericutes																									
Thaumarchaeota																									
Thermodesulfobacteria																									
Thermotogae																									
Verrucomicrobia																									

Table S4.2. Distribution of CRISPR leader sequences according to subtype and host taxonomy. Each subset of leader sequences was subjected to Uclust search with 0.6 threshold. This filtration resulted in removal of leader sequences that have 60% or higher similarity to avoid prediction of motifs biased to specific organism. It was assumed that conserved motifs found in remaining sequences should provide homology between leader sequences, while the regions outside the motifs are diverse.

Class	Type	Subtype	CRISPRDetect_Edit annotation with cas-gene proximity and Leader-motif-assisted classification
CLASS 1	Type I	IA	509
		IB	3036
		IC	2856
		ID	506
		IE	9914
		IF	4807
		IF3	96
		IU	356
	Type III	IIIA	2167
		IIIB	2771
		IIIC	77
		IIID	928
		IIIF	4
	Type IV	IVA	12
		IVB	2
CLASS 2	Type II	IIA	1266
		IIB	42
		IIC	2172
	Type V	VA	46
		VB	7
		VE	89
		VF	19
		VJ	70
		VK	4
	Type VI	VIA	15
		VIB1	115
		VIB2	2
		VIC	11

Table S4.3. Subtype distribution of CRISPR arrays predicted from the "Relaxed" search with CRISPRDetect and annotation with repeat-, *cas*- and leader-based methods.

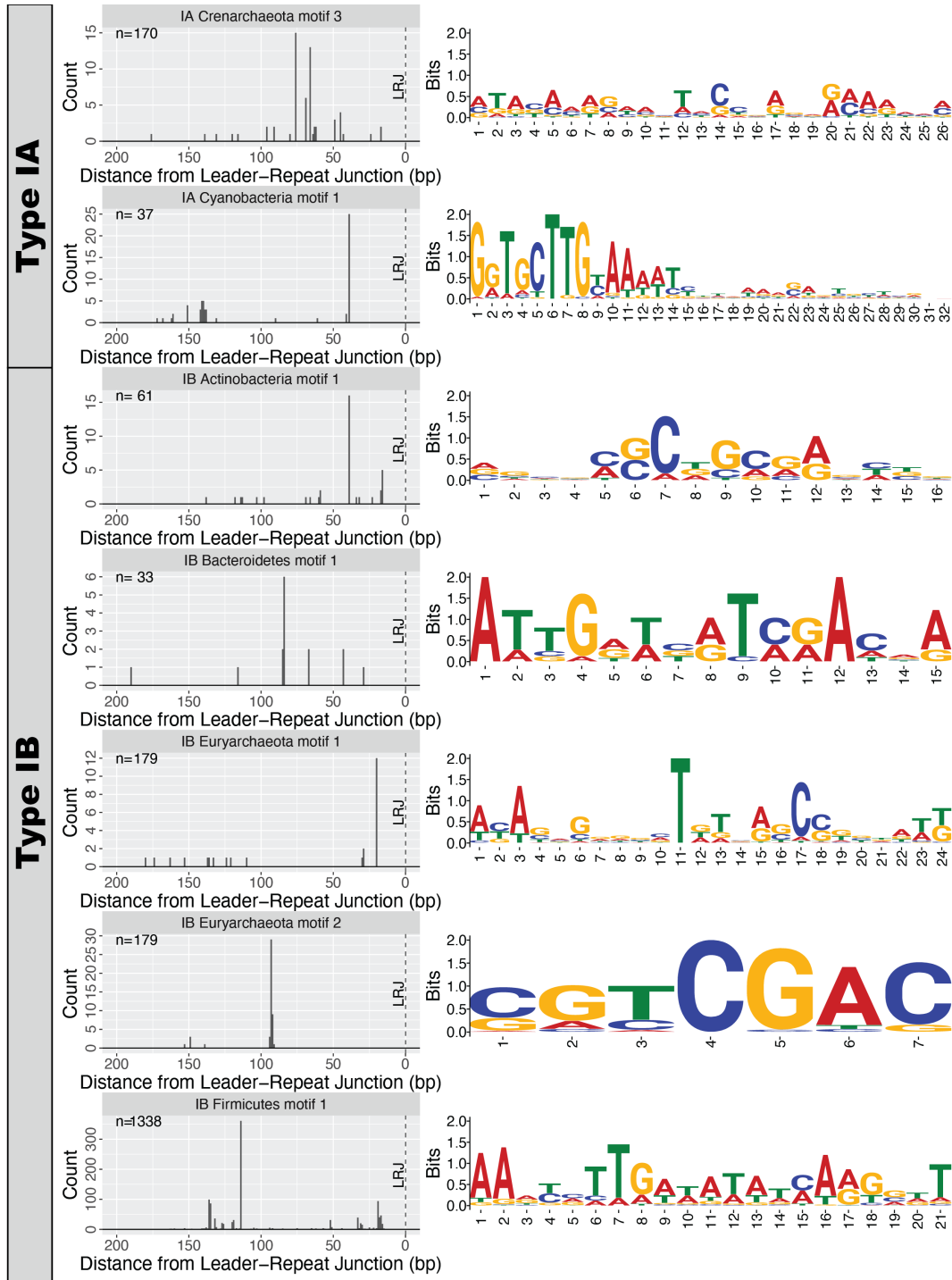


Figure S4.1. Each predicted motif possesses narrow distribution in the group CRISPR leader sequences that they originally identified.

Tally of the midpoints of motifs identified in CRISPR leaders belonging to different subtypes. Position weight matrices (PWMs) are shown for subtype and taxonomy-specific motifs found in CRISPR leaders.

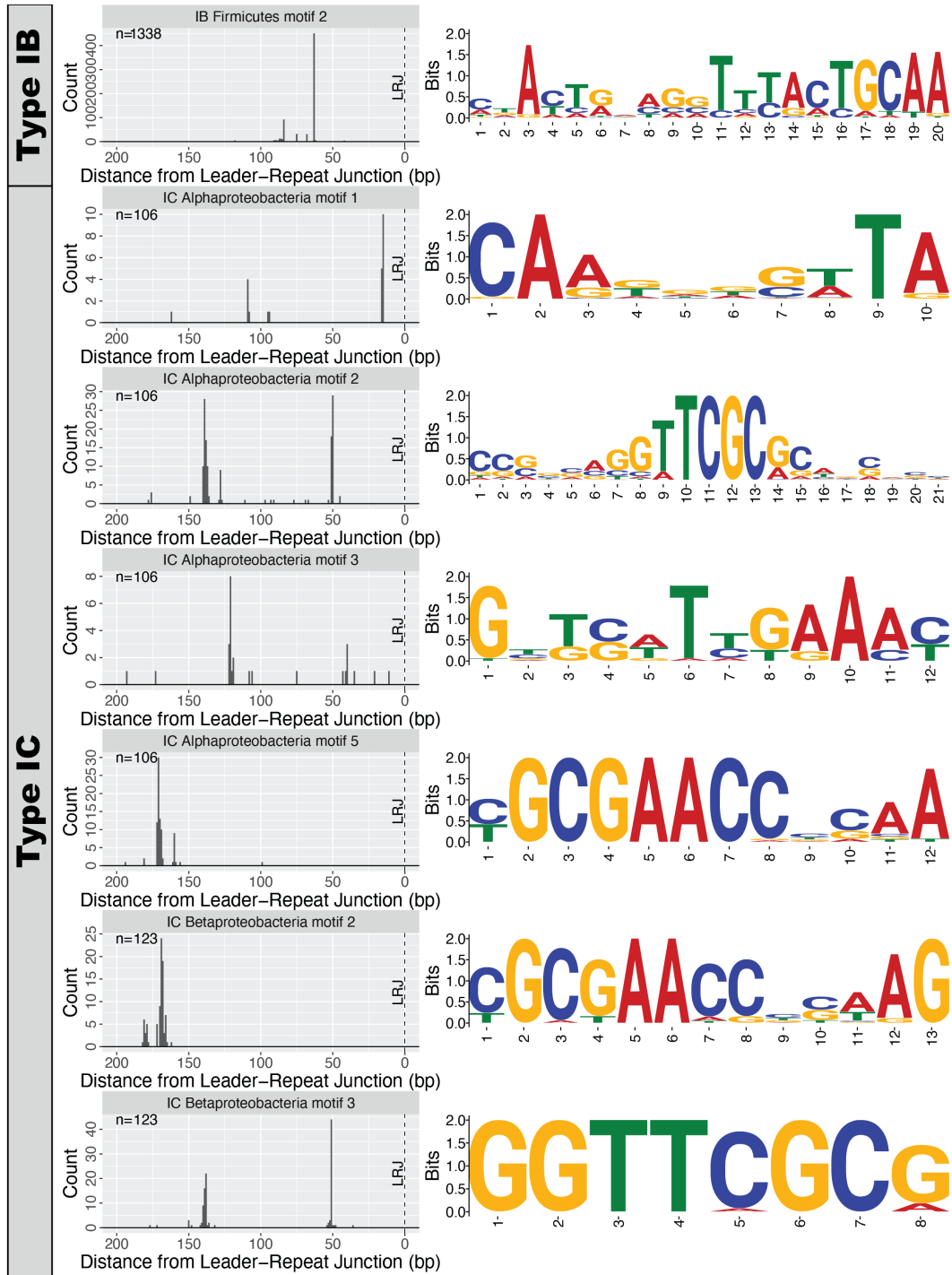


Figure S4.1 Continued.

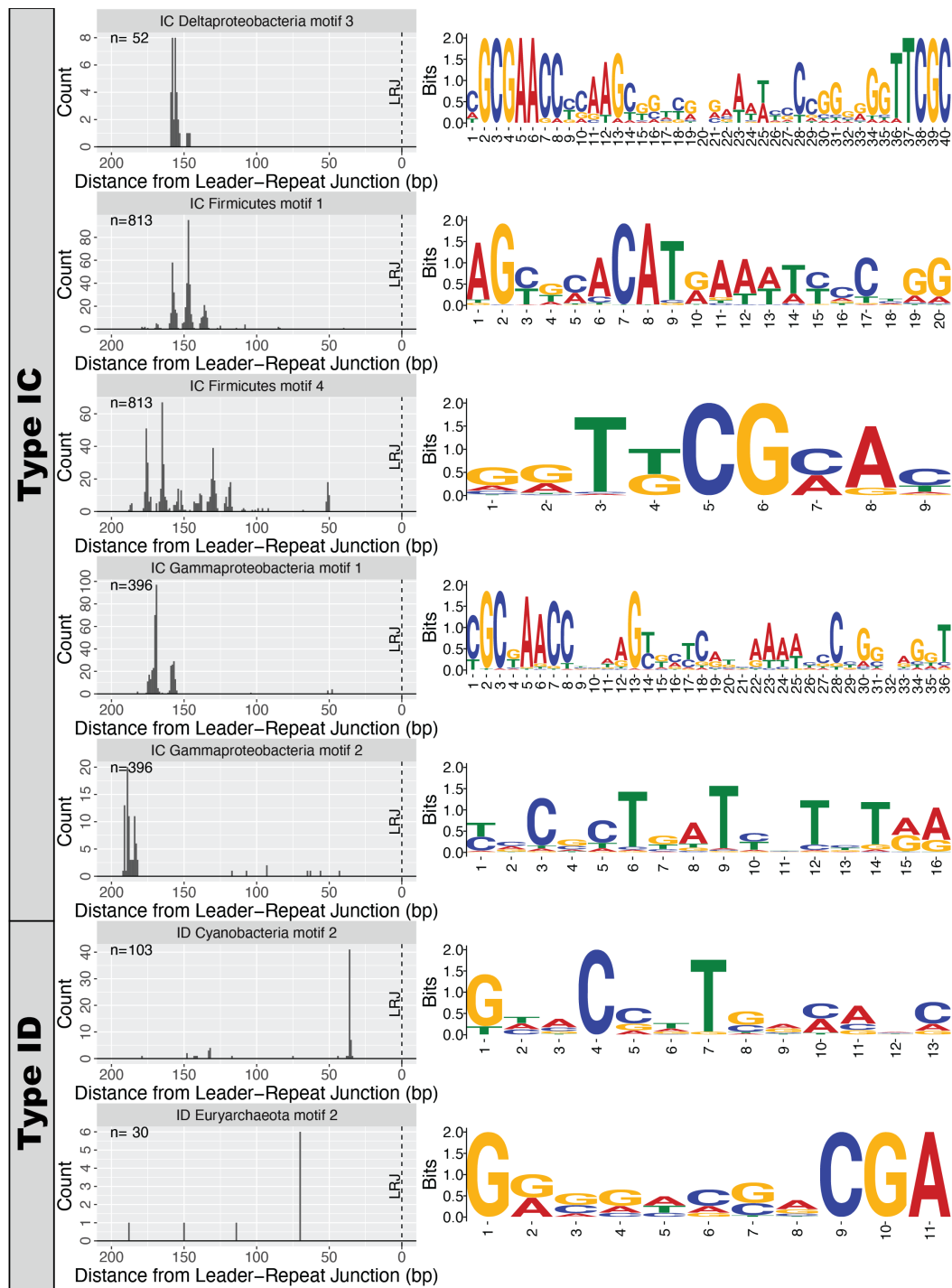


Figure S4.1 Continued.

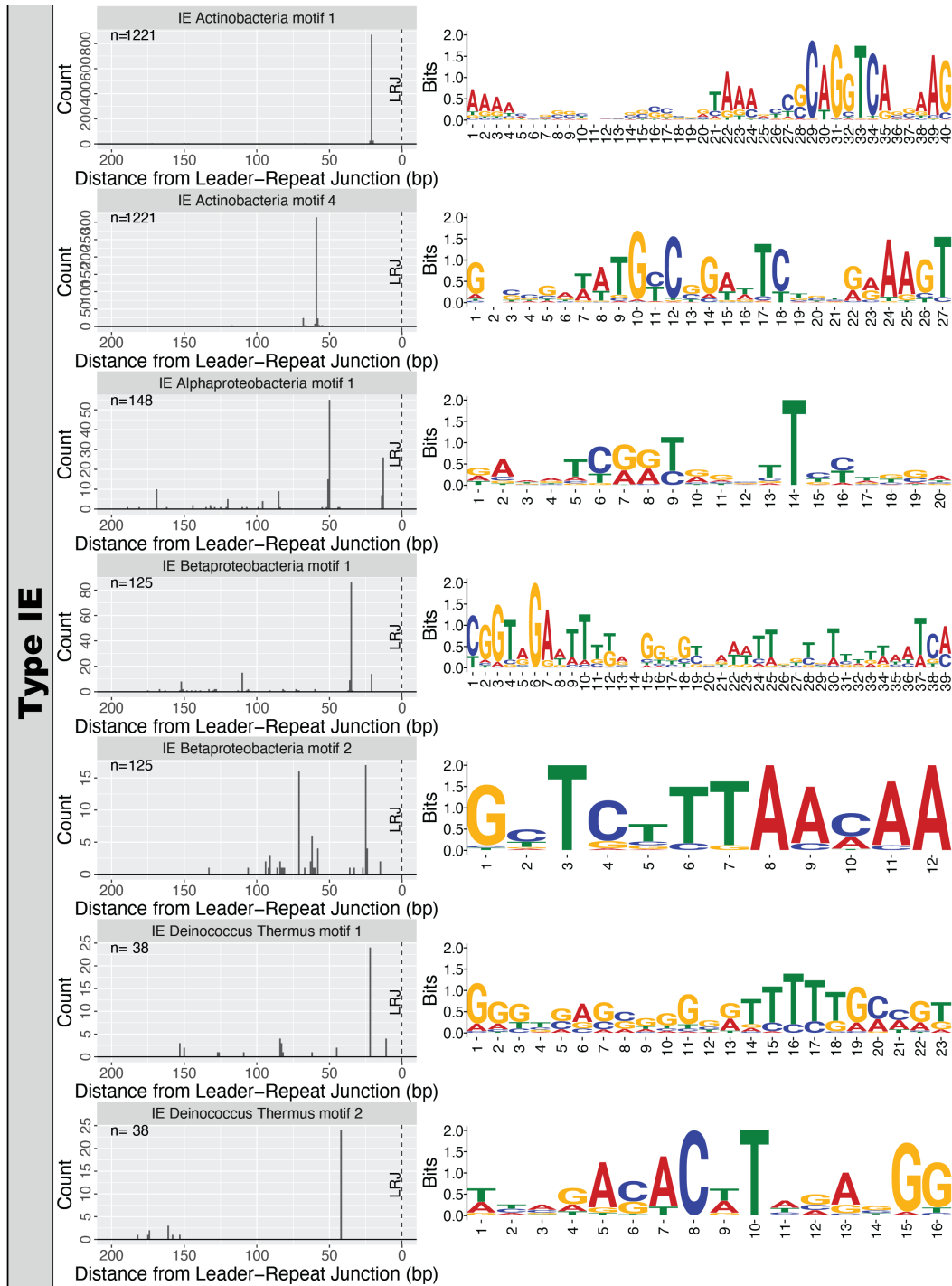


Figure S4.1 Continued.

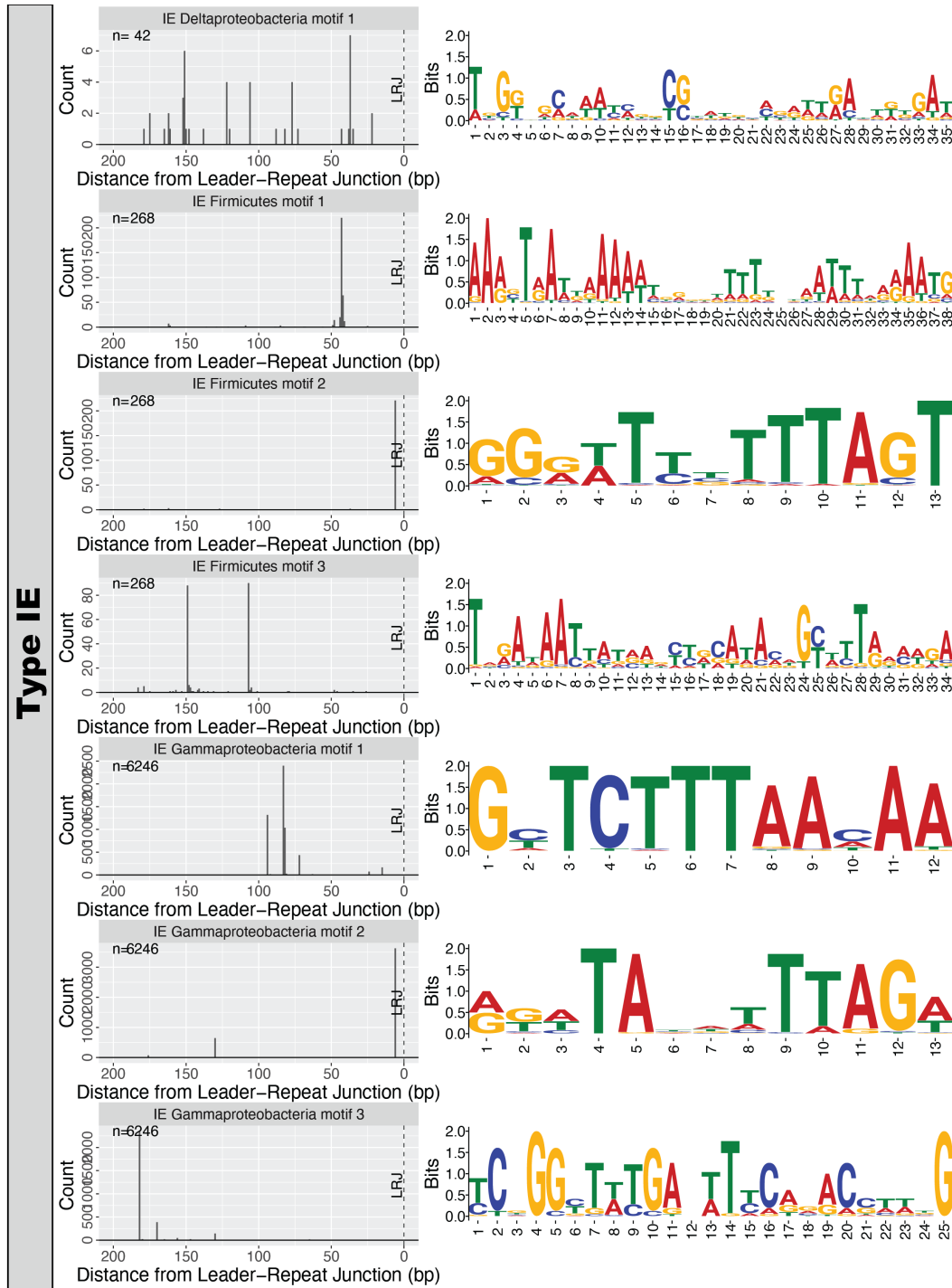


Figure S4.1 Continued.

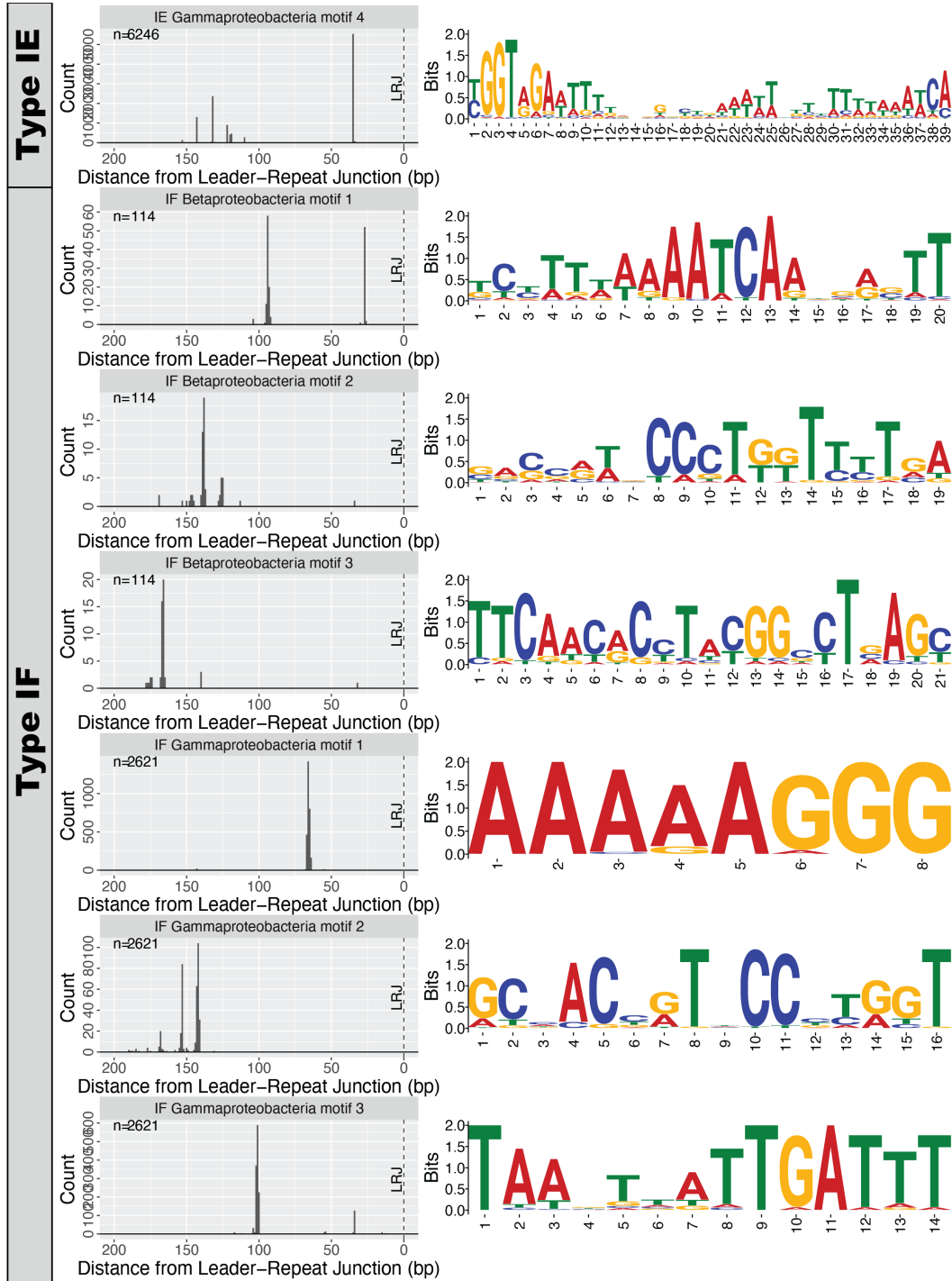


Figure S4.1 Continued.

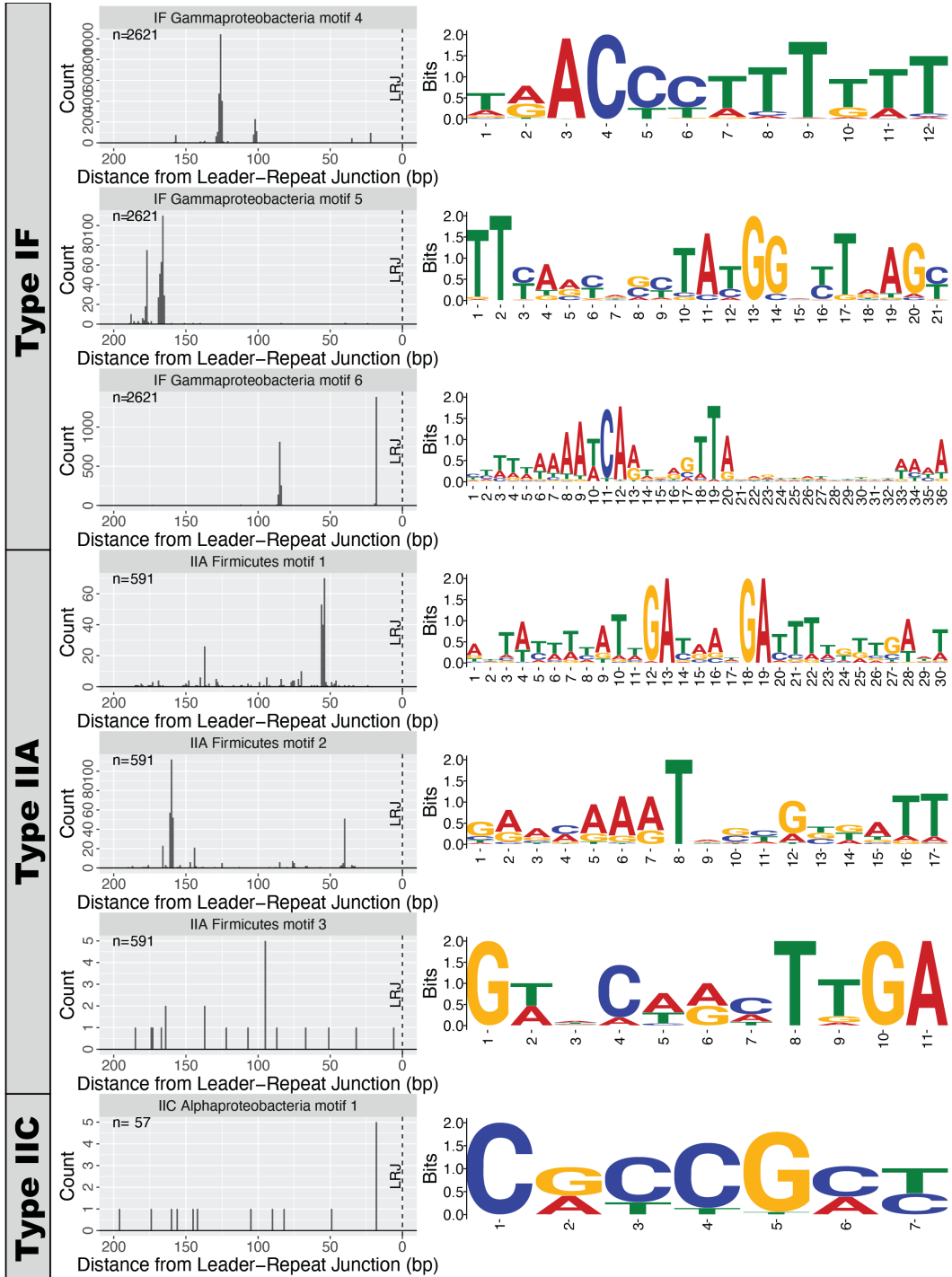


Figure S4.1 Continued.

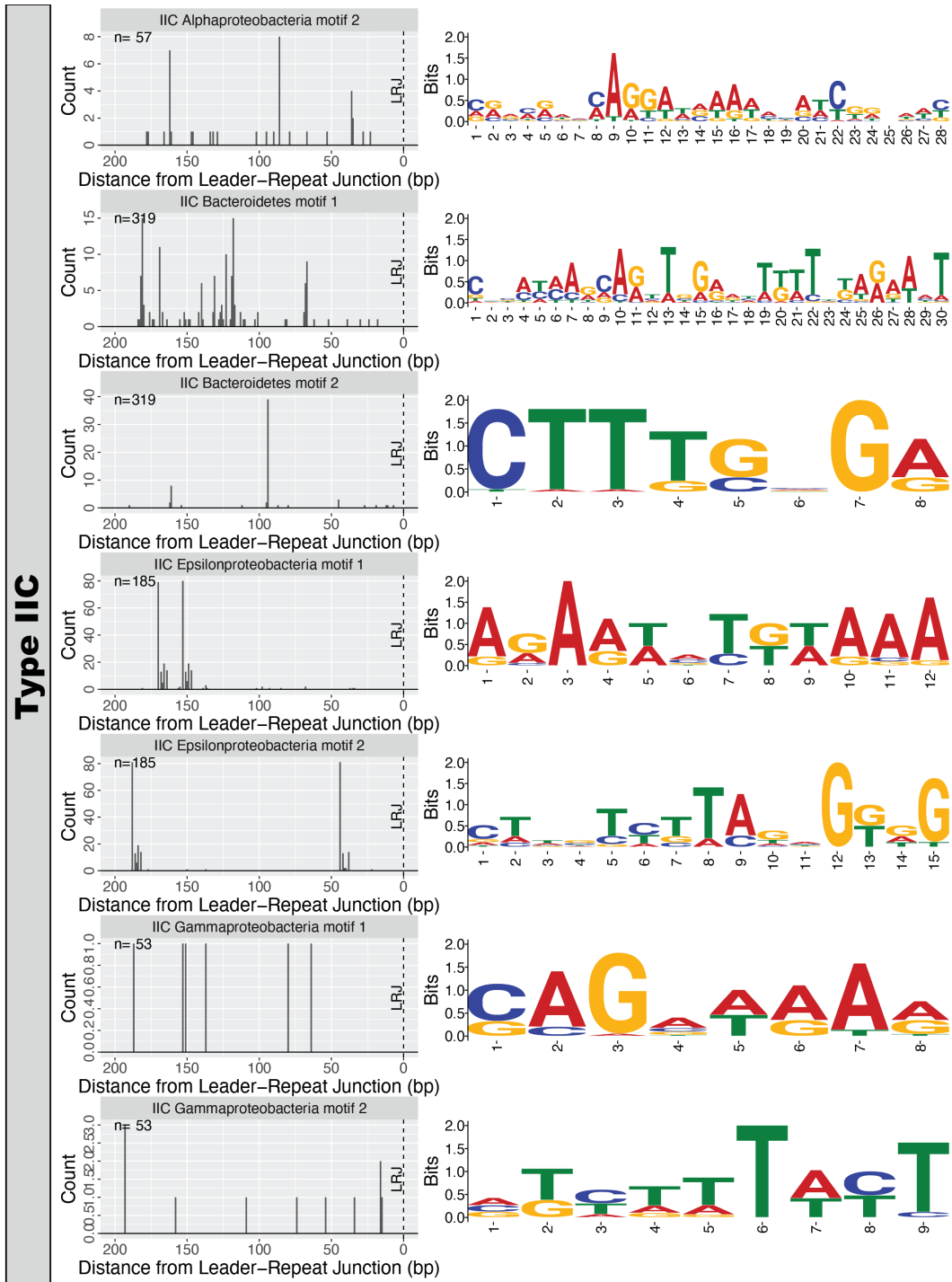


Figure S4.1 Continued.

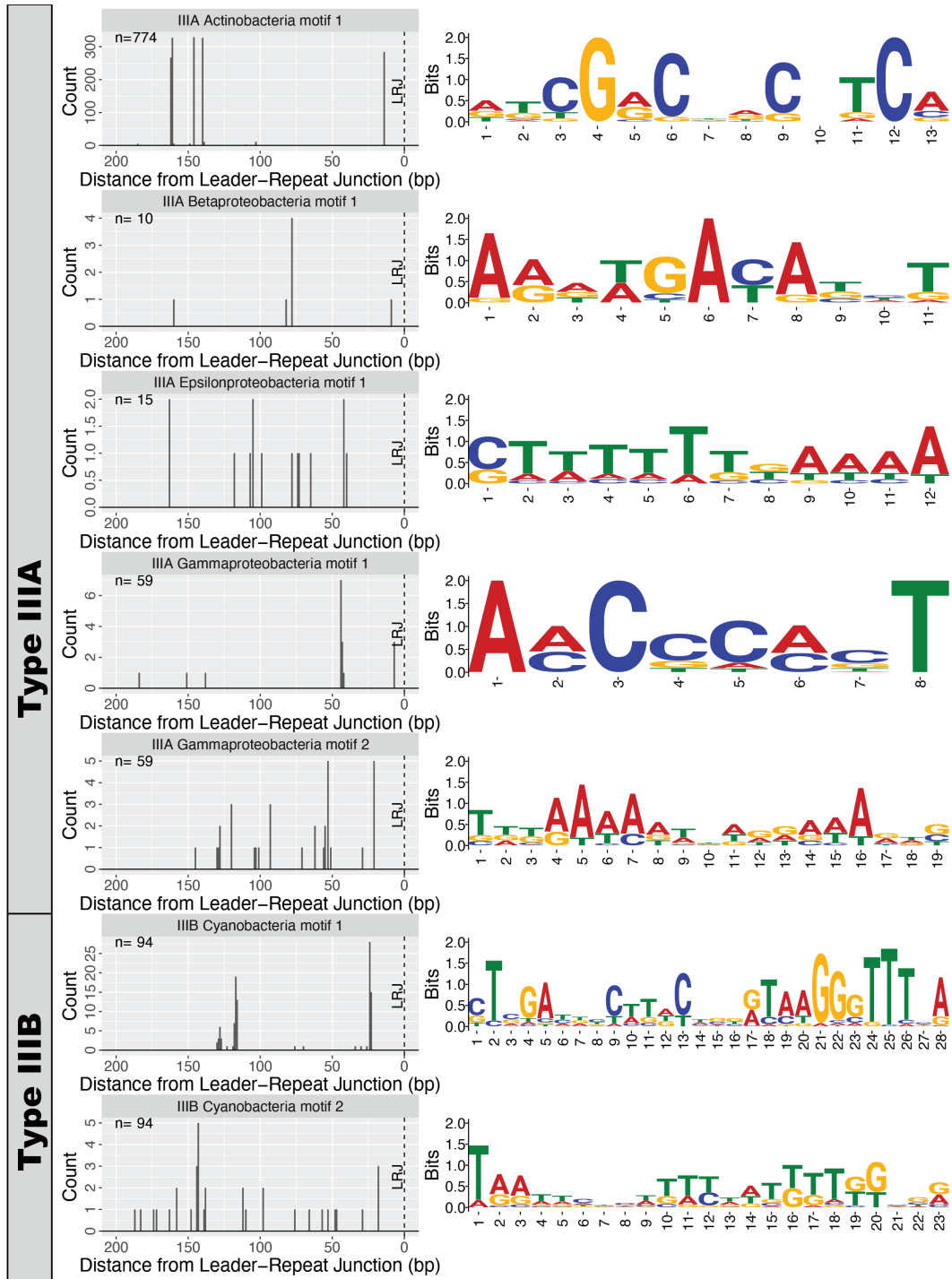


Figure S4.1 Continued.

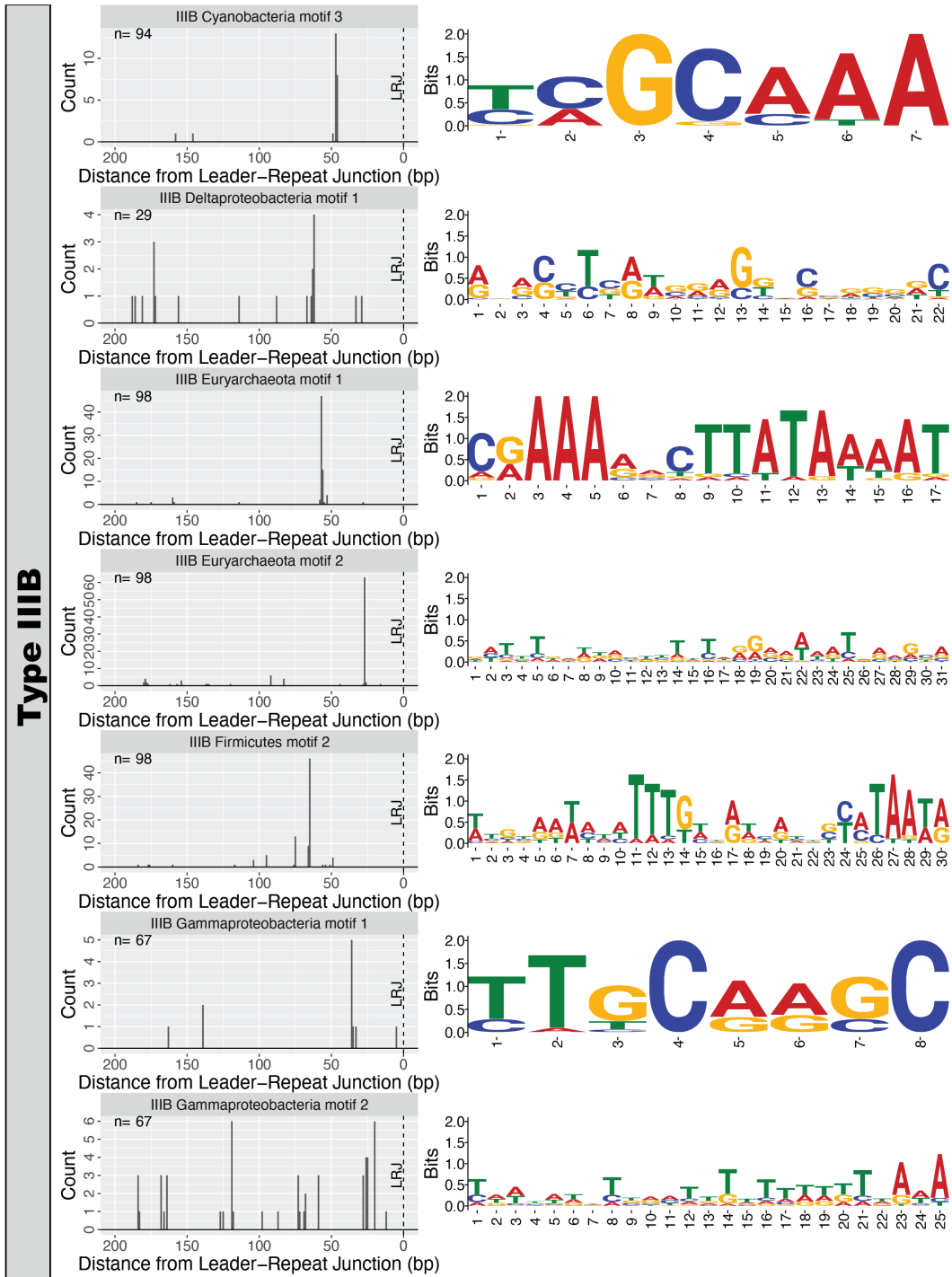


Figure S4.1 Continued.

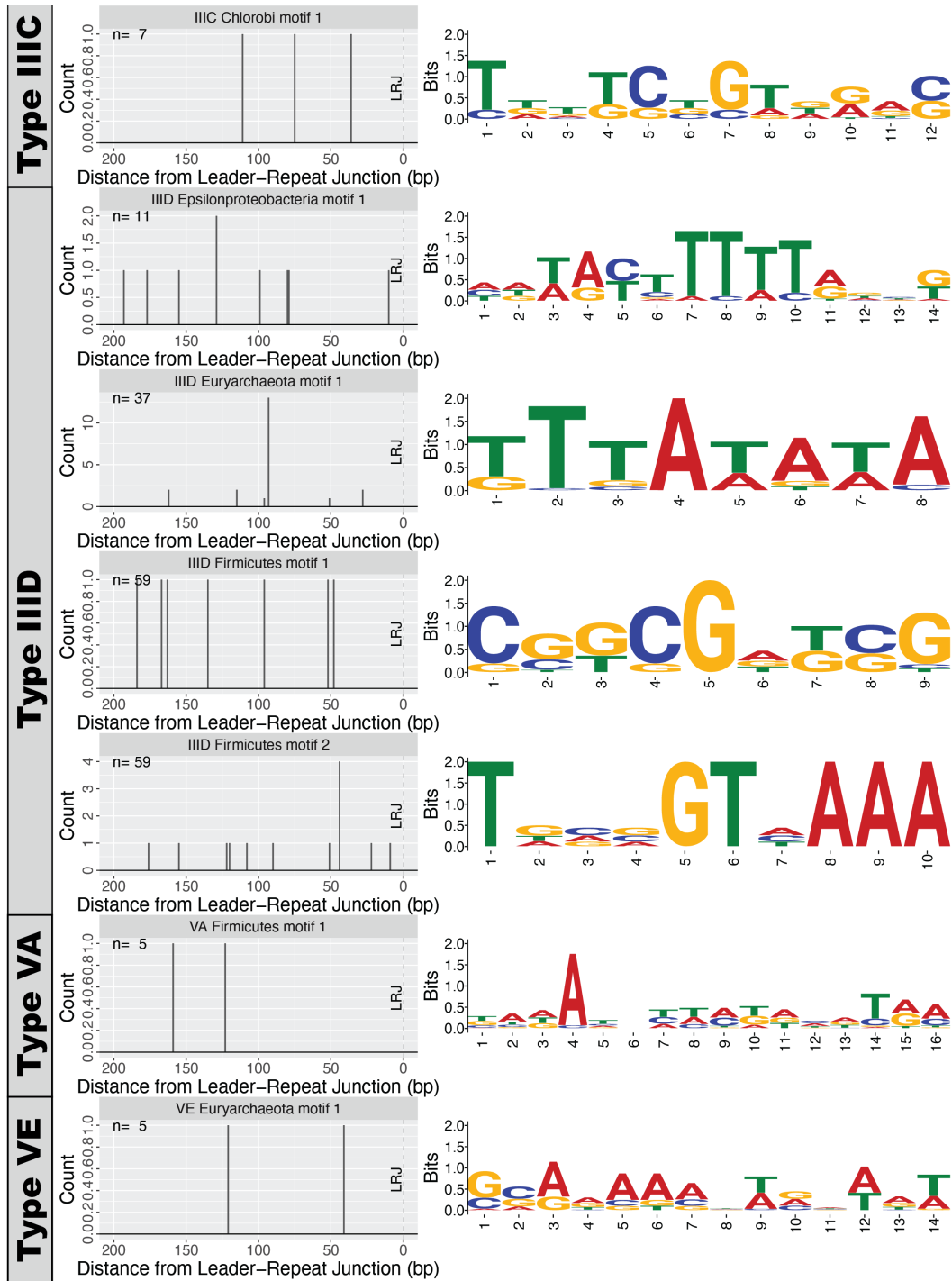
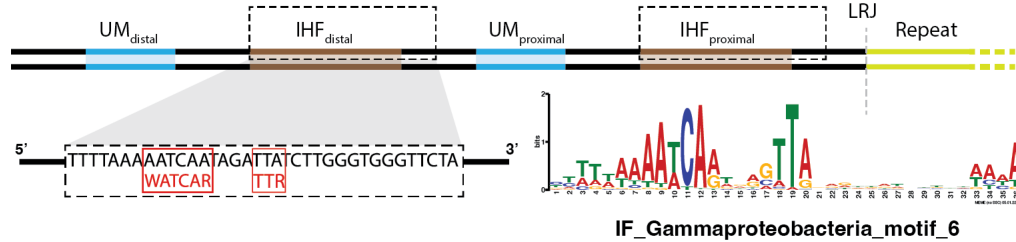


Figure S4.1 Continued.

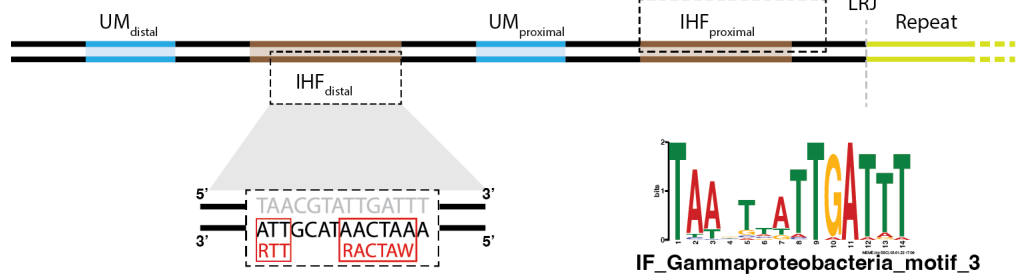
a) Sense strand PWM:

IF CRISPR leader from *Halotalea alkalilenta* strain IHB B 13600
(NZ_CP015243.1, Position: 275,952-276,179bp)



Anti-sense strand PWM:

IF CRISPR leader from *Yersinia pseudotuberculosis* strain FDAARGOS_583
(NZ_CP033709.1, Position: 4,047,478-4,047,705bp)



b) IF_Betaproteobacteria_motif_1: PWM represents both proximal and distal occurrences of IHF.



IE_Gammaproteobacteria_motif_4: PWM represents both proximal and distal occurrences of I-E_UM + IHF.

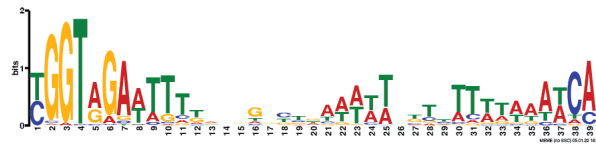


Figure S4.2. PWMs predicted from de novo motif prediction analysis that represents different occurrences of IHF binding site in different subtypes.

(a) Two examples of IHF_{distal} strand preference represented by two different PWMs specifically distinguish IHF sites observed in sense or anti-sense strands of I-F CRISPR leaders. **(b)** Representative PWMs of IHF binding site overserved in CRISPR leaders from different subtype/host-taxonomy pairs.

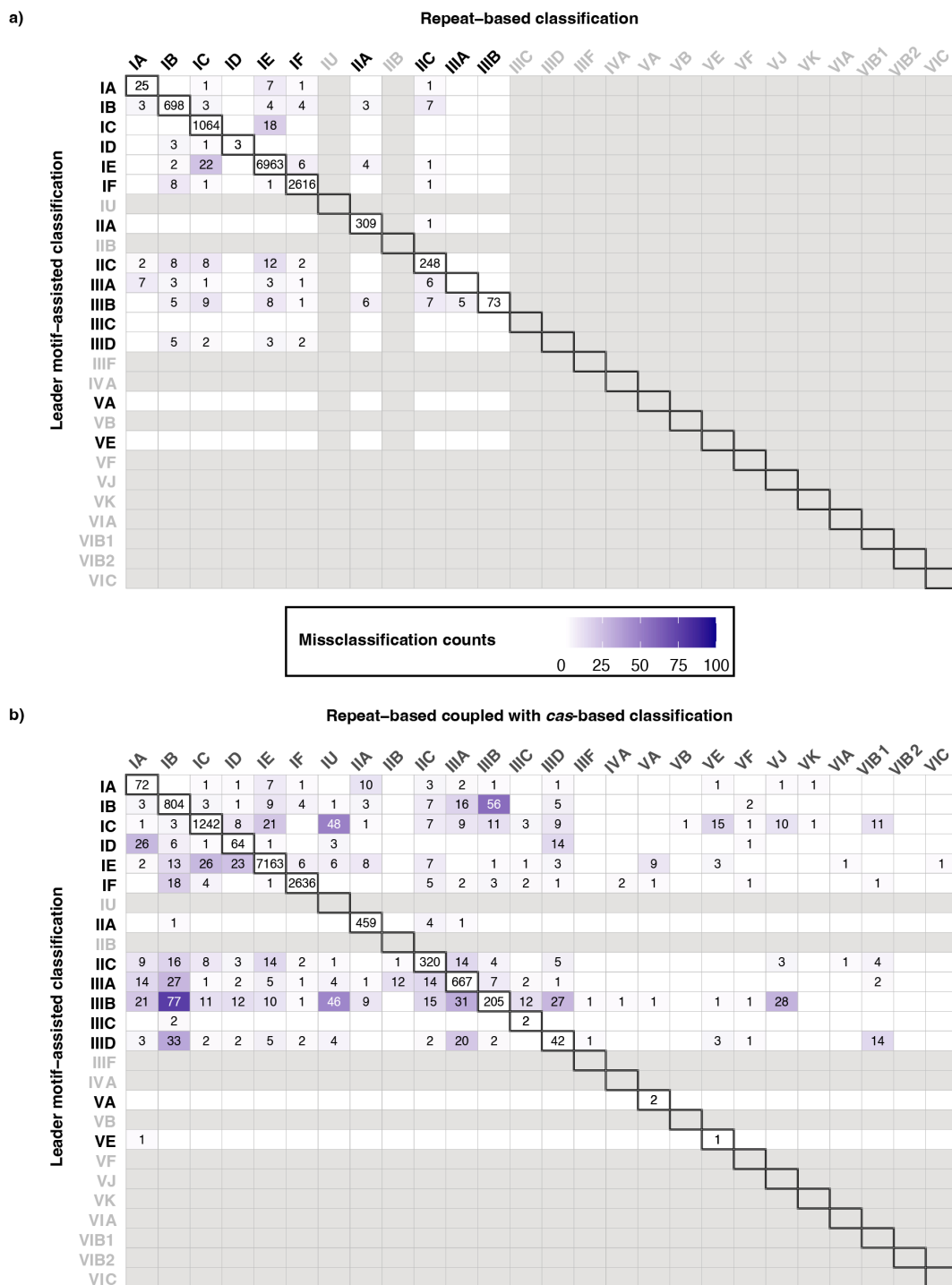


Figure S4.3. Comparison of the motif-assisted array predictions against the high-confidence predictions generated by traditional classification methods.

(a) Each motif used for array prediction are searched with the CRISPR leader sequences that only subjected to classification by CRISPRDetect (repeat-based subtype annotation). Possible

misclassifications of leader sequences are highlighted with blue according to count of array predictions observed in different subtype datasets. Compared to misclassification results from repeat and *cas*-based method, leader-motif assisted annotation perform equivalent to repeat-based method with few misclassifications. Rows and columns are greyed out in cases where the motif data or leader sequence data are unavailable. **b)** Identified subtype-specific motifs were searched within the CRISPR leader sequences that previously classified by CRISPRDetect coupled with *cas*-based array detection. HMM profiles of *cas* genes are used to search within the 20kb upstream and downstream of CRISPR arrays in addition to CRISPRDetect annotation. Misclassified leaders through leader-motif-assisted method are represented in each subtype and highlighted in blue according to count of array predictions observed.

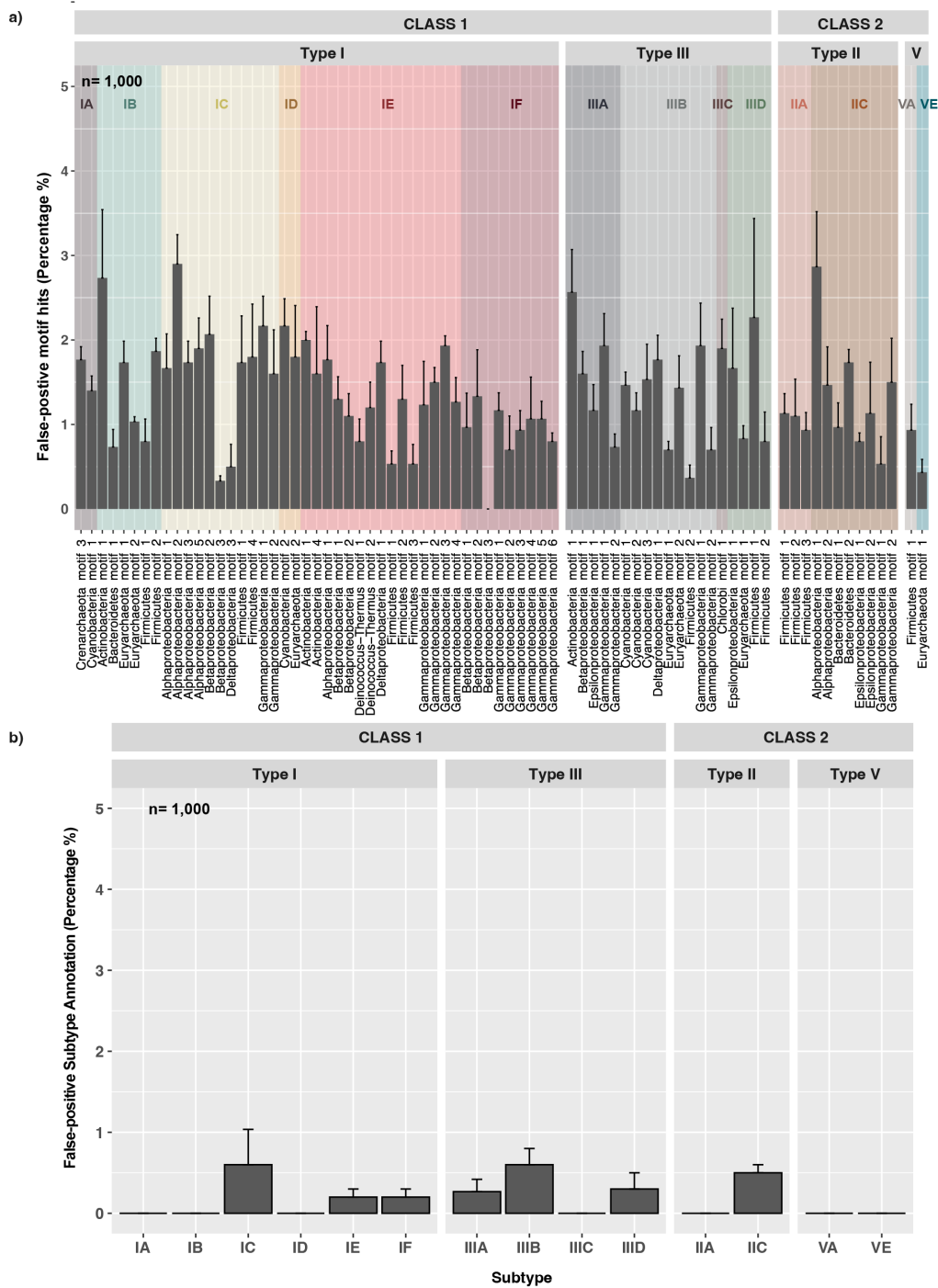


Figure S4.4. Leader Motif-assisted Annotation in Non-leader Sequences yields less than 0.6% false-annotation.

Overview of motif-assisted array annotation results in a subset of 1,000 randomly generated 200bp long sequences. Analysis performed in technical triplicates. **(a)** Identification of 77 previously predicted CRISPR leader motifs showed less than 4% error rate due to significant false-positive

sequence matches. **(b)** Discrete positioning requirement for each motif has applied to test whether false-positive matches contribute to labeling arrays as CRISPR. Among the all false-positive hits, only 21 motifs had significant distribution score that included in the false-array annotation. The overall false annotation rate was less than 0.6% in all subtype specific CRISPR leader callings.

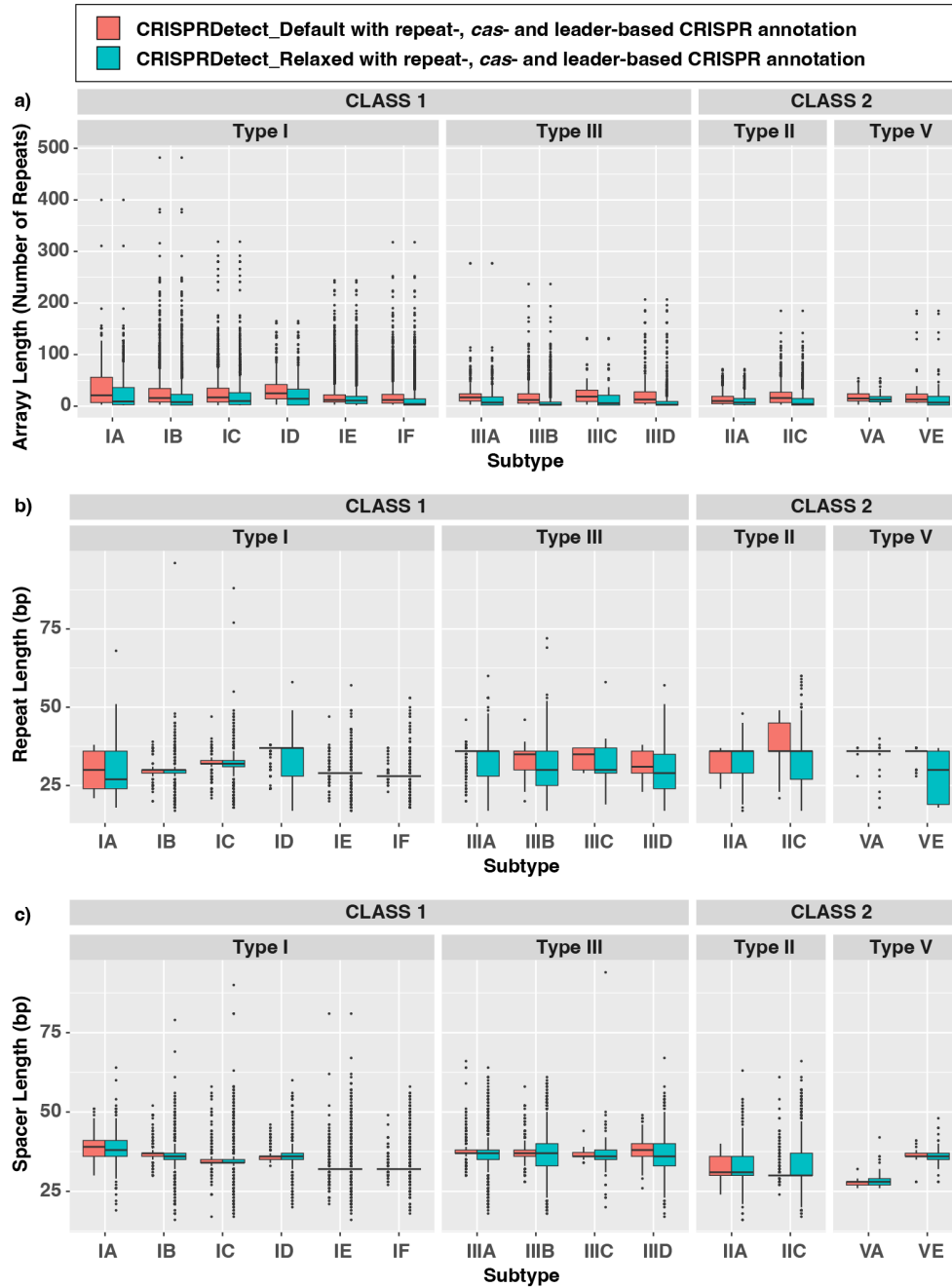


Figure S4.5. CRISPR Subtype-specific distribution of array, repeat and spacer lengths predicted from edited version of CRISPRDetect coupled with leader motif-assisted annotation.

Box plot of **a)** array lengths, **b)** repeat lengths, and **c)** spacer lengths for subtype categorized CRISPRs. Edited CRISPRDetect search coupled with leader-motif-assisted annotation captures shorter CRISPR arrays that previously filtered by the CRISPRDetect search with the default parameters.

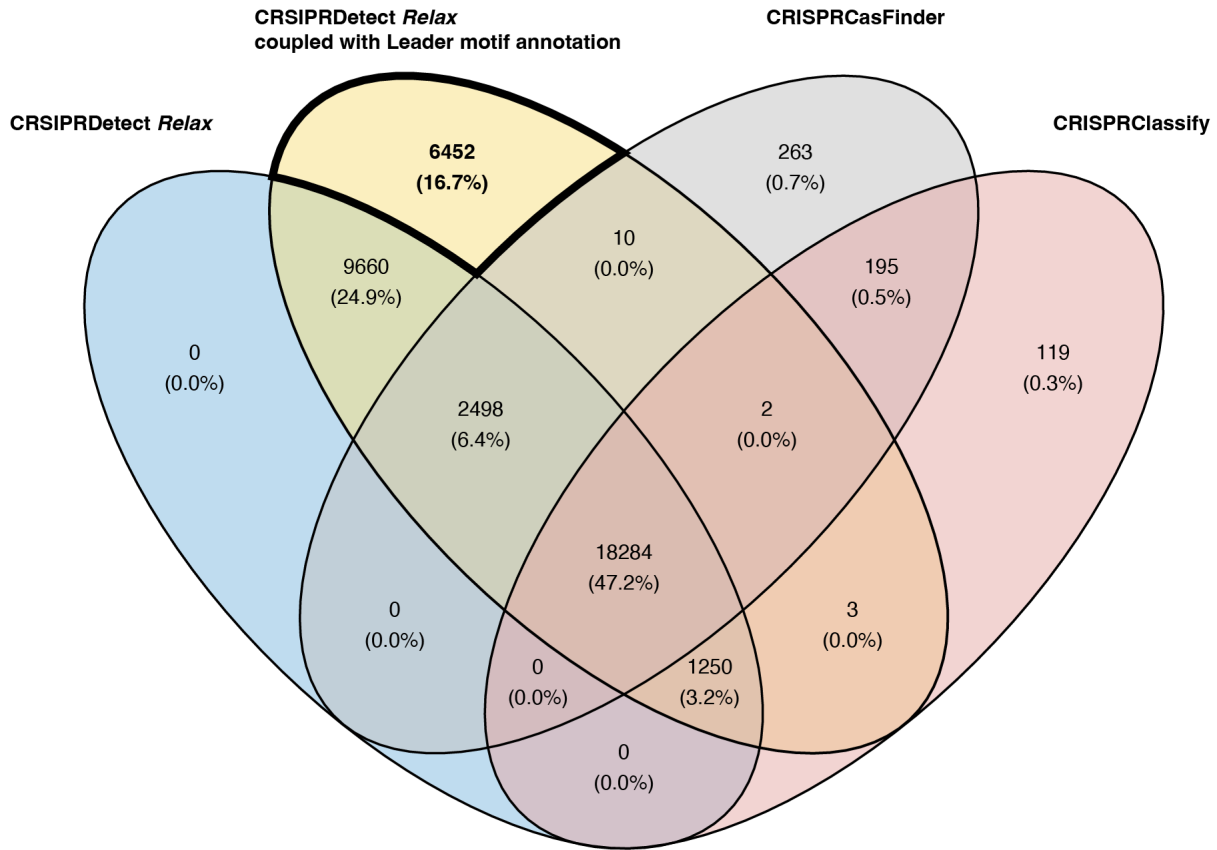


Figure S4.6. The number of high-confidence CRISPR loci increases with motif-assisted CRISPR search.

Leader-motif-assisted CRSIPRDetect recall level in comparison with other tools on the annotated arrays dataset. This dataset was compiled out of 25,502 bacterial and 398 archaeal complete genomes obtained from NCBI (July 2021). Only 47.2% of the CRISPR arrays (18,284 arrays) are predicted by all tools. 16.7% of arrays (6,452 arrays) are only predicted by the edited version of CRSIPRDetect search that subjected to leader-motif-assisted annotation.

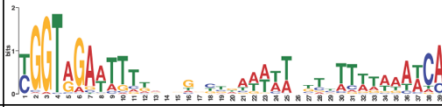
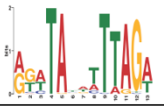
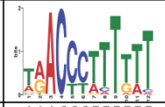
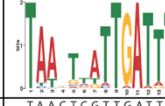
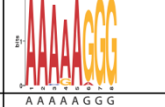
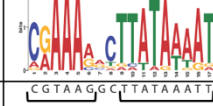
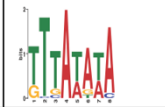
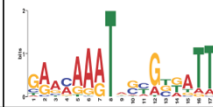
Mechanism	Motif ID	PWM/Matching motif sequence	Motif ID	PWM/Matching motif sequence	Leader accession
Adaptation	IR-proximal & IHF-proximal (I-E Gammaproteobacteria motif 4)	 Sequence: T G G T A G A T T G T T G A T G T G G A A T A A A T T T G T T T A A A A C A G A T Motif: IR-proximal IHF-proximal	I-E Gammaproteobacteria motif 2	 Sequence: A T G T A T G C T T A G T	NZ_CP028305.1_Array_2 Position: 55-1
	IR-distal (I-F Gammaproteobacteria motif 4)	 Sequence: A A A C C C T T T T T			NZ_CP024240.1_Array_4 Position: 122-121
	IHF-distal (I-F Gammaproteobacteria motif 3)	 Sequence: ... T A A C T C G T T G A T T T ...			NZ_CP024240.1_Array_4 Position: 108-95
	IR-proximal (I-F Gammaproteobacteria motif 1)	 Sequence: A A A A A G G G			NZ_CP024240.1_Array_4 Position: 69-62
	IHF-proximal (I-F Gammaproteobacteria motif 6)	 Sequence: ... G C C T T A A T A A C A A T A A G A T A G C G C T G T T T T C T T C C T			NZ_CP024240.1_Array_4 Position: 36-1
	Histone binding site (III-B Euryarchaeota motif 2)	 Sequence: G A T A T A G T T A C T C T G T A G G A G T A T T G G G G C A			NC_018092.1_Array_1 Position: 43-13
Transcription	BRE & TATA (III-B Euryarchaeota motif 1)	 Sequence: C G T A A G G C C T T A T A A A T T Motif: BRE TATA			NC_018092.1_Array_1 Position: 65-49
	TATA (III-D Euryarchaeota motif 1)	 Sequence: T T T A T A A A ...			NZ_CP009516.1_Array_3 Position: 97-90
	Stem-loop structure (II-A Firmicutes motif 2)	 Sequence: ... A G A C A A A T A G T G C G A T T ...			NZ_L5483394.1_Array_1 Position: 49-33

Table S4.4. Detected Leader motifs with previously known functions.

The table presents identified motifs in the CRISPR leader sequences that have been previously associated with specific biological functions including CRISPR adaptation and transcription machineries. Each row displays a position weight matrix (PWM) representation of the leader motifs, accompanied by a representative example sequence, accession of the leader sequence and the position of the demonstrated motif relative to the leader-repeat junction (LRJ). In instances where a single PWM captures two distinct leader motifs, both motifs are illustrated in the same row, with each motif distinctly bracketed. The associated biological functions, whether related to CRISPR adaptation or transcription of the CRISPR loci, are specified in the adjacent column.

CHAPTER FIVE

CLARIFYING CRISPR: WHY REPEATS IDENTIFIED IN THE HUMAN GENOME
SHOULD NOT BE CONSIDERED CRISPRS

Contribution of Authors and Co-Authors

Manuscript in Chapter 5

Author: Murat Buyukyoruk

Contributions: Designed the research, performed the research, prepared manuscript figures, wrote the manuscript.

Co-Author: William S. Henriques

Contributions: Designed the research, prepared manuscript figures, performed the research.

Co-Author: Blake Wiedenheft

Contributions: Designed the research, wrote the manuscript.

Manuscript Information

Murat Buyukyoruk, William S. Henriques, Blake Wiedenheft

The CRISPR Journal

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☒ Published in a peer-reviewed journal

Published by Mary Ann Liebert, Inc.

Volume 5

Number 3

10 April 2023

DOI: 10.1089/crispr.2022.0106

CLARIFYING CRISPR: WHY REPEATS IDENTIFIED IN THE HUMAN GENOME
SHOULD NOT BE CONSIDERED CRISPRS

Murat Buyukyoruk¹, William S. Henriques¹, Blake Wiedenheft¹

¹Department of Microbiology and Immunology, Montana State University, Bozeman, MT
59717, USA

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPR) and their associated genes (*cas*) are essential components of adaptive immune systems that protects bacteria and archaea from viral infection. CRISPR-Cas systems are found in about 40% of bacterial and 85% of archaeal genomes, but not in eukaryotic genomes. Recently, a paper published in *Communications Biology* reported the identification of 12,572 putative CRISPRs in the human genome, which they call “hCRISPR”. Here, we attempt to reproduce this analysis and show that repetitive elements identified as putative CRISPR loci in the human genome contain neither the repeat-spacer-repeat architecture nor the *cas* genes characteristic of functional CRISPR systems.

Introduction

Repetitive DNA sequences are diverse and abundant in a broad range of species, from bacteria to mammals, where they participate in many different biological functions¹. CRISPRs belong to a specific family of DNA repeats found in bacteria and archaea that all share a common architecture^{2, 3}. Each CRISPR locus consists of a series of short repeat sequences (20-50 bps) separated by unique spacer sequences of a similar length (**Figure 5.1A**)⁴. The repeat sequences within a CRISPR locus are generally conserved but repeats between different CRISPR loci can vary in both sequence and length². The diversity of these elements makes them challenging to confidently identify, especially when the number of repeat sequences within a locus is small.

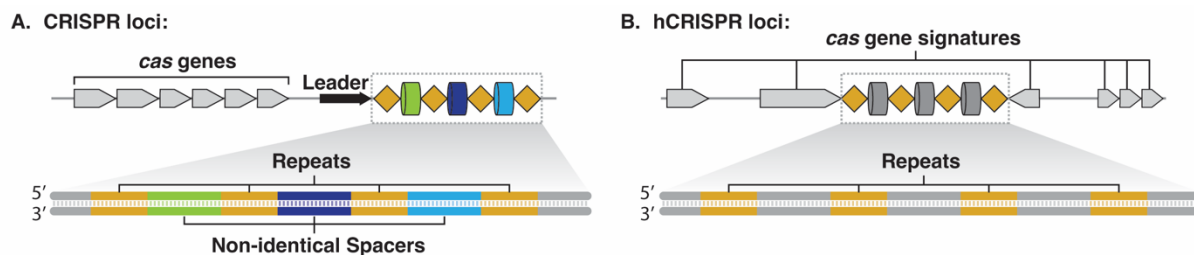


Figure 5.1. Comparison of architectural features in CRISPRs and hCRISPRs.

A) Structure of typical CRISPR-Cas locus in bacteria and archaea. **B)** Structure of a recently reported “hCRISPR”.

The discovery of CRISPR-based adaptive immunity and the ongoing adaptation of these systems for applications in genome editing has engaged scientists from nearly every discipline and many perspectives. Thus, the recent paper by van Riet *et al.*, which reported the discovery of more than 12,500 CRISPR loci in the human reference genome (GRCh38.p13; GCF_000001405.39), comes as a surprise⁵. They report that hCRISPRs are similar to prokaryotic CRISPRs in terms of

array architecture, repeat length, repeat sequence, and that hCRISPR are flanked by CRISPR-associated (*cas*) genes (**Figure 5.1B**)⁵. Moreover, they claim that hCRISPRs are found across all human chromosomes and that the distribution is directly correlated to chromosome length. Collectively, these claims indicate that CRISPR loci are preserved in the human genome and their association with *cas* genes implies a functional connection that will have profound biological and biotechnological implications. Here, we attempt to reproduce this analysis and show that repetitive elements identified in the human genome contain neither the repeat-spacer-repeat architecture nor the *cas* genes characteristic of functional CRISPR systems.

Methods

CRISPR Predictions

The human reference genome (GRCh38.p13; GCF_000001405.39) was downloaded from the NCBI RefSeq Assembly database (193 sequences) (August 22nd, 2022)⁶. Local versions of CRISPRCasFinder⁷ (v4.2.20; CasFinder v2.0.3), CRISPRDetect⁸ (v3), and CRISPRCasTyper⁹ (v1.6.4) were used to identify putative CRISPR arrays in the downloaded human reference genome. We used the following search command for CRISPRCasFinder “*-minSP 21 -maxSP 72 -fast*”, and “*-array_quality_score_cutoff 0 -minimum_word_repeatation 2 -minimum_no_of_repeats 2*” for CRISPRDetect. Default parameters were used for the CRISPRCasTyper. Predicted putative arrays from CRISPRCasFinder, CRISPRDetect, and CRISPRCasTyper were sorted based on the assigned confidence score. Evidence-level scores 3 or 4 by CRISPRCasFinder, array quality score above 4 by CRISPRDetect, and probability score above 0.75 by CRISPRCasTyper were considered as high-confidence arrays. Results from each CRISPR detection tool were compared to identify shared hits. Twenty candidate arrays predicted

by CRISPRDetect were removed from the list of high-confidence arrays due to low confidence scores for these same arrays by CRISPRCasFinder. The remaining 72 arrays were then considered candidate arrays. Two of these arrays with the highest confidence scores from each of the three detection methods were manually inspected.

In addition, the T2T assembly of the human genome with the Y chromosome (GCF_009914755.1 T2T-CHM13v2.0) was downloaded from the NCBI RefSeq Assembly database (January 10th, 2023)¹⁰. CRISPRDetect was used to predict putative CRISPR arrays in the T2T assembly of the human genome using the same search parameters described above. Predicted putative arrays by CRISPRDetect were sorted by quality score.

Genomes of *Aedes aegypti* (GCF_002204515.2)¹¹, *Arabidopsis thaliana* (GCF_000001735.4)¹², *Caenorhabditis elegans* (GCF_000002985.6)¹³ were downloaded from the NCBI RefSeq Assembly database (August 22nd, 2022). CRISPRDetect was used to predict putative CRISPR arrays in the downloaded genomes with the same search parameters described above. Arrays identified by CRISPRDetect were sorted by quality score.

The mitochondria genomes of *Vicia faba* (X59246.1)¹⁴ and the assembly of *Acanthamoeba polyphaga* mimivirus genome (GCF_000888735.1)¹⁵ were downloaded from the NCBI RefSeq database (January 12th, 2023). Previously reported CRISPR-like sequences from each genome were extracted and manually inspected.

Identification of *cas* Genes in the Human Genome

CRISPRCasFinder and CRISPRCasTyper were used to identify *cas* genes in the human reference genome. CRISPRCasFinder failed to predict any putative *cas* gene in the human reference genome. Putative *cas* genes predicted by CRISPRCasTyper were subjected to a filtering

step according to E-value ($< 10^{-5}$) and sequence coverage cutoff (sequence coverage $> 50\%$). In addition, translated CDS sequences of both the human reference genome ($n=60,090$) and the T2T assembly of the human genome ($n=65,591$) were downloaded from the NCBI RefSeq Assembly database^{6, 10}. One hundred and twenty one *cas* gene HMM profiles from CRISPRCasdb were used to query the CDS sequences using MacsyFinder (v1.0.5; HMMER 3.2.1)^{16, 17}. Queries were performed using the following parameters: “*macsyfinder --sequence-db <open_reading_frames> --db-type gembase -d <CRISPR_subtype_definitions> -p <HMM_profiles> -w 50 -vv all*”. Results from the HMMER analysis were filtered with the same E-value threshold and sequence coverage criteria described above.

Comparison of Repeat Length Distribution of CRISPRs from Prokaryotes and hCRISPRs

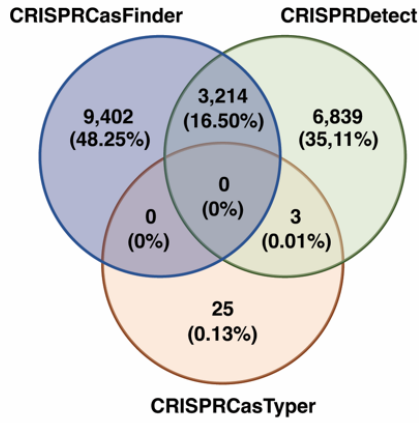
The repeat sequences from 19,483 unique putative hCRISPR arrays were extracted using a Python v2.7 script that uses the Bio SeqIO package¹⁸. In addition, the repeats of bacterial and archaeal CRISPRs were downloaded from CRISPRCasdb ($n=28,712$). The distribution of repeat lengths was visualized in RStudio by comparing the list of repeats from CRISPRs and hCRISPRs using ggplot¹⁹.

Results and Discussion

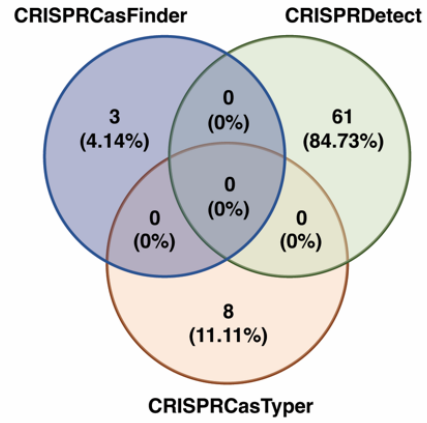
To understand how the putative CRISPR arrays reported by van Riet *et al.*, may have been missed by the rest of the scientific community and to learn what these new CRISPRs might teach us about the evolution and function of CRISPR systems, we searched the human reference genome (GRCh38.p13; GCF_000001405.39) for CRISPRs using a combination of CRISPRCasFinder⁷, CRISPRDetect⁸ and CRISPRCasTyper⁹ according to the search parameters defined by the

authors⁵. Collectively, we identified a total of 19,483 unique putative arrays in the human reference genome.

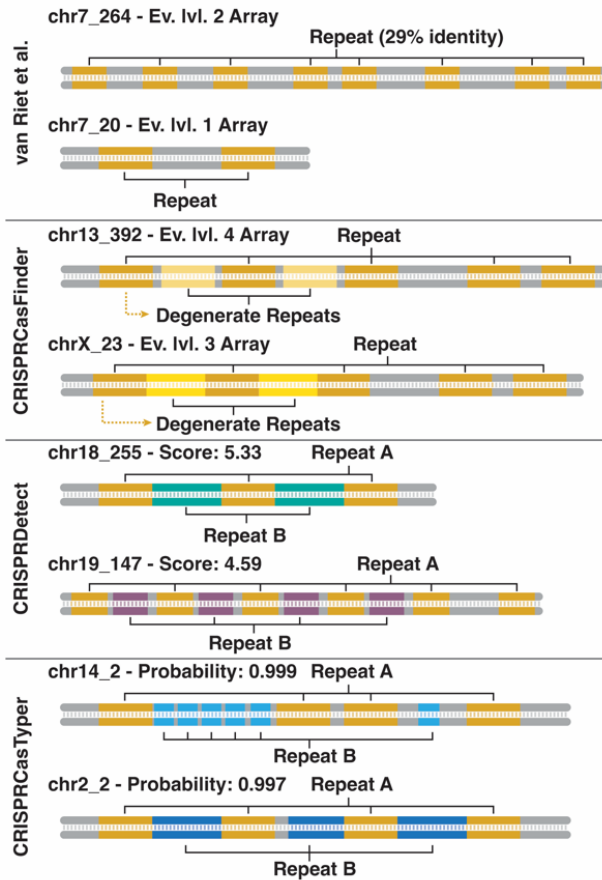
A. All putative arrays (n=19,483)



B. High confidence arrays (n=72)



C.



D.

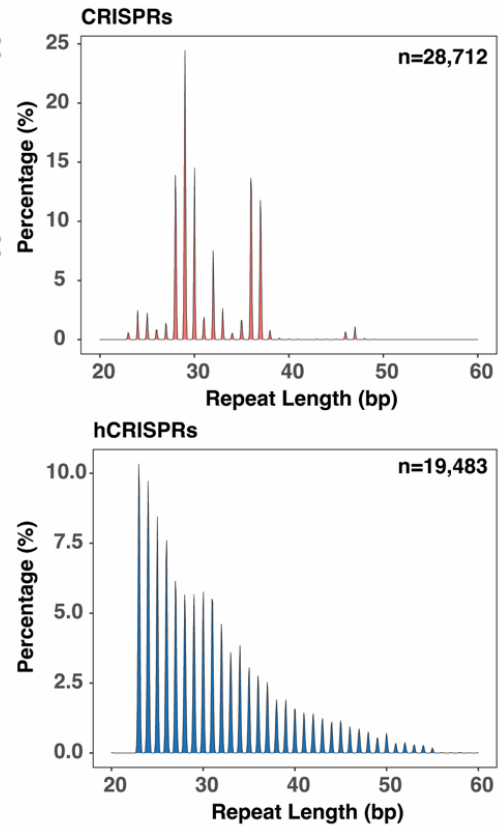


Figure 5.2. Repetitive elements in the human genome do not contain architectural features consistent with CRISPR loci.

A) The human reference genome was queried using CRISPRCasFinder, CRISPRDetect and CRISPRCasTyper. A total of 19,483 unique putative arrays were predicted. 3,214 arrays were detected by both CRISPRCasFinder and CRISPRDetect, whereas only three of the arrays were predicted by both CRISPRDetect and CRISPRCasTyper. None of the arrays predicted by CRISPRCasFinder and CRISPRCasTyper were the same. **B)** CRISPRCasFinder predicted only three arrays with high confidence (evidence-level 3 or 4), while CRISPRDetect predicted 61 candidate arrays (array quality score > 4) and CRISPRCasTyper assigned eight arrays as high confidence (probability > 75%). None of these predicted arrays are shared between all three CRISPR detection methods. **C)** Architecture of two arrays presented by van Riet *et al.* (top), followed by the top two most confident predictions from each of the three detection methods. The long array identified by van Riet *et al.* contains repeats that are only 29% identical and spacers that range in size. The other array is too short to predict with confidence. The top two arrays predicted by CRISPRCasFinder contain spacer sequences that are degenerate versions of the repeat. The other four arrays identified by CRISPRDetect and CRISPRCasTyper contain spacers that are related, resulting in arrays with tandem repeats. **D)** The direct repeats of bacterial and archaeal CRISPRs compared to lengths of the repeats in hCRISPR. All repeat sequences were extracted and the number of each repeat corresponding to a particular length is plotted for bacterial and archaeal genomes (top, n=28,712) and the human reference genome (bottom, n=19,483).

Results from CRISPRCasFinder (n=12,616) and CRISPRCasTyper (n=28) were nearly identical to those reported by van Riet *et al.*, while CRISPRDetect (n=10,056) identified 6,403 additional arrays (**Figure 5.2A**). The exact search parameter used for CRISPRDetect were not provided by the authors (i.e., `-array_quality_score_cutoff 0 -minimum_word_repetition 2 -minimum_no_of_repeats 2`), so the additional CRISPR loci identified here are anticipated to reflect differences in the search criteria.

CRISPR prediction tools assign confidence scores based on a combination of parameters. These parameters include the number of repeats in an array, conservation of repeat sequences, and the presence of non-identical spacer sequences^{7-9, 20}. CRISPRDetect incorporates additional factors, such as the identification of previously described repeat sequences and detection of conserved *cas* genes (i.e., *cas1* or *cas2*), which are used to calculate confidence scores. While confidence scores are not mentioned by van Riet *et al.*⁵, we sorted the arrays according to scores

calculated by each algorithm. CRISPRCasFinder identified a total of three arrays with high confidence scores, CRISPRDetect identified 61 arrays with high confidence, and CRISPRCasTyper assigned eight arrays with high confidence (**Figure 5.2B**).

To further evaluate these predictions, we manually inspected the two examples presented by van Riet *et al.*⁵ (neither of which earned a high confidence score by any of the search methods), as well as two arrays with the highest confidence scores from each of the three detection methods (**Figure 5.2C**). The two hCRISPR examples presented by van Riet *et al.* include one long and one short locus (**Figure 5.2C, top**)⁵. The long array contains eight “repeat” sequences that are 29% identical to each other and are separated by unique spacers sequences that vary in length (8-75 bps) (**Figure 5.2**). The short “hCRISPR” contains two identical 28 bp repeats and a single 39 bp spacer sequence. We cannot definitively eliminate the possibility that the short array (composed of two repeats) is a CRISPR, but no *cas* genes are found flanking the CRISPR, the repeats do not correspond to any of repeats in CRISPRCasdb¹⁶, and the spacer sequence only matches to itself. In addition to the two examples provided by van Riet *et al.*, we selected two arrays with high confidence scores from each of the three different CRISPR detection algorithms (**Figure 5.2C, middle**). High confidence arrays identified by CRISPRCasFinder contain “spacer” sequences that are degenerate versions of the repeat sequences, whereas putative arrays identified using CRISPRDetect and CRISPRCasTyper all contain “spacer” sequences that are 60% to 72% identical to one another. Thus, these arrays contain an alternating series of two repeats (i.e., Repeat A and Repeat B), rather than repeats separated by unique spacer sequences (**Figure 5.2C, bottom**).

To determine if the repeat sequences in the putative human CRISPR loci are similar in length to those typically found in prokaryotic CRISPRs, we downloaded the direct repeats of

bacterial and archaeal CRISPRs from CRISPRCasdb¹⁶ (n=28,712) and compared them to the distribution of repeat lengths from putative CRISPR loci in the human genome (**Figure 5.2D**). CRISPR repeats identified in bacterial and archaeal genomes are primarily 27-28 bp or 36-37 bp in length, while repeats identified in the human reference genome are primarily 23 bp and decrease in frequency as the repeat length increases. While the average length of the repeat sequences is similar (30 bp as identified by van Riet *et al.*), the distributions are very different.

In addition to CRISPR detection, the authors also report the detection of *cas* genes in the flanking regions of 336 hCRISPRs⁵. However, neither the sequences for these ORFs nor significance values for these *cas* genes are provided. Our efforts to reproduce the identification of *cas* genes in the human reference genome using CRISPRCasFinder and CRISPRCasTyper failed. CRISPRCasFinder failed to predict any putative *cas* gene, while CRISPRCasTyper failed to predict significant hits (E-value < 10⁻⁵ and sequence coverage > 50%). To further verify these results, we downloaded 121 profile Hidden Markov Models (HMMs) for *cas* genes from CRISPRCasdb¹⁶ and queried the human genome using MacSyFinder (v1.0.5)²¹. Consistent with CRISPRCasFinder and CRISPRCasTyper, we find no significant hits (E-value < 10⁻⁵ and sequence coverage > 50%) to *cas* genes in the human genome.

The human reference genome (GRCh38.p13) screened by van Riet *et al.* was originally assembled nearly 20 years ago⁶. While early drafts of the human genome provided critical new insight into human biology, these assemblies preferential excluded many repetitive sequences. However, the recently released Telomere-to-Telomere (T2T) assembly of the human genome, has resolved this problem¹⁰. Due to the repetitive nature of the human genome, we hypothesized the T2T genome might contain CRISPRs not found in earlier versions. To test this hypothesis, we

searched the T2T genome (GCF_009914755.1 T2T-CHM13v2.0) using the methods described above. Collectively, we identified a total of 10,481 putative CRISPR arrays in the T2T genome. CRISPRDetect identified 78 putative CRISPRs with high confidence (52 arrays were identical to those predicted in human reference genome); however, none of these arrays maintain the architectural characteristics of a CRISPR.

Repetitive sequences make up ~50% of the human genome^{10,22}. This is not a unique feature of the human genome, but characteristic of many eukaryotic genomes, where repetitive elements make up 1-80% of genomes²³⁻²⁵. Several annotation tools have been developed to classify eukaryotic repetitive elements and notably, none have found evidence for CRISPRs^{26,27}. Using the same methods described here for finding CRISPRs in prokaryotic genomes, we found 3,498 putative CRISPRs in the *Aedes aegypti* genome (310 Mb), 1,419 putative CRISPRs in the *Arabidopsis thaliana* genome (30 Mb), 490 putative CRISPRs in the *Caenorhabditis elegans* genome (15 Mb). In this small, non-representative sample of genomes, the number of putative CRISPRs identified correlates with the size of the genome, which is similar to the distribution of putative CRISPRs identified by van Riet *et al.* (“hCRISPRs are found across all chromosomes and their distribution is directly correlated to the chromosome length”).

Apart from the work by van Riet *et al.*, CRISPR loci have been previously reported in the mitochondrial genome of a plant (*Vicia faba*) and the genome of a giant mimivirus that infects single-celled eukaryotes^{3,15}. However, neither of these repeats meet the CRISPR criteria described above (**Figure S5.1**)^{2,28}.

Overall, the evidence for functional CRISPR-Cas systems in the human genome and other eukaryotic genomes remains unsupported. While the phylogenetic and functional diversity of

CRISPRs continue to expand, (to date) CRISPR-Cas adaptive immune systems remain restricted to prokaryotes.

Acknowledgements

We thank members of the Wiedenheft laboratory for comments, suggestions, and critical feedback. Research in the Wiedenheft laboratory is supported by National Institutes of Health (R35GM134867), the M.J. Murdock Charitable Trust, a young investigator award from Amgen, a generous gift from the Rosolowsky family, and the Montana State University Agricultural Experimental Station (USDA NIFA).

References

1. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions*. Nat Rev Genet, 2011. **13**(1): p. 36-46.
2. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes*. Molecular Microbiology, 2002. **43**(6): p. 1565-1575.
3. Mojica, F.J., et al., *Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria*. Mol Microbiol, 2000. **36**(1): p. 244-6.
4. Ishino, Y., M. Krupovic, and P. Forterre, *History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology*. J Bacteriol, 2018. **200**(7).
5. van Riet, J., et al., *CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue*. Commun Biol, 2022. **5**(1): p. 338.
6. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly*. Genome Res, 2017. **27**(5): p. 849-864.
7. Couvin, D., et al., *CRISPRCasFinder; an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins*. Nucleic Acids Res, 2018. **46**(W1): p. W246-W251.
8. Biswas, A., et al., *CRISPRDetect: A flexible algorithm to define CRISPR arrays*. BMC Genomics, 2016. **17**: p. 356.
9. Russel, J., et al., *CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci*. CRISPR J, 2020. **3**(6): p. 462-469.
10. Nurk, S., et al., *The complete sequence of a human genome*. Science, 2022. **376**(6588): p. 44-53.
11. Matthews, B.J., et al., *Improved reference genome of Aedes aegypti informs arbovirus vector control*. Nature, 2018. **563**(7732): p. 501-507.
12. Pucker, B., et al., *A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set*. PLoS One, 2019. **14**(5): p. e0216233.
13. Consortium, C.e.S., *Genome sequence of the nematode C. elegans: a platform for investigating biology*. Science, 1998. **282**(5396): p. 2012-8.

14. Flamand, M.C., et al., *Sequence and Transcription Analysis of Mitochondrial Plasmids Isolated from Cytoplasmic Male-Sterile Lines of Vicia-Faba*. Plant Molecular Biology, 1992. **19**(6): p. 913-923.
15. Levasseur, A., et al., *MIMIVIRE is a defence system in mimivirus that confers resistance to virophage*. Nature, 2016. **531**(7593): p. 249-52.
16. Pourcel, C., et al., *CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers*. Nucleic Acids Res, 2020. **48**(D1): p. D535-D544.
17. Prakash, A., et al., *The HMMER Web Server for Protein Sequence Similarity Search*. Curr Protoc Bioinformatics, 2017. **60**: p. 3 15 1-3 15 23.
18. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics*. Bioinformatics, 2009. **25**(11): p. 1422-1423.
19. Villanueva, R.A.M. and Z.J. Chen, *ggplot2: Elegant Graphics for Data Analysis, 2nd edition*. Measurement-Interdisciplinary Research and Perspectives, 2019. **17**(3): p. 160-167.
20. Alkhnbashi, O.S., et al., *CRISPR-Cas bioinformatics*. Methods, 2020. **172**: p. 3-11.
21. Abby, S.S., et al., *MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems*. PLoS One, 2014. **9**(10): p. e110726.
22. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-51.
23. Goemann, C.L.C., et al., *Genome sequence, phylogenetic analysis, and structure-based annotation reveal metabolic potential of Chlorella sp. SLA-04*. Algal Research, 2023. **69**: p. 102943.
24. Hufford, M.B., et al., *De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes*. Science, 2021. **373**(6555): p. 655-662.
25. Udriste, A.A., et al., *Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the Ve Locus*. Int J Mol Sci, 2022. **23**(17).
26. Bao, W., K.K. Kojima, and O. Kohany, *Rebase Update, a database of repetitive elements in eukaryotic genomes*. Mob DNA, 2015. **6**: p. 11.
27. Hubley, R., et al., *The Dfam database of repetitive DNA families*. Nucleic Acids Res, 2016. **44**(D1): p. D81-9.
28. Mohanraju, P., et al., *Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems*. Science, 2016. **353**(6299): p. aad5147.

Supplemental Information

Manually inspected example arrays presented by van Riet *et al.* and the top two arrays from each of the three detection methods, including CRISPRCasFinder, CRISPRDetect and CRISPRCasTyper.

- **van Riet *et al.***

>chr7_264 | NC_000007_264 | Start: 56832024 | Stop: 56832500 | Evidence Level: 2 | Num. of predicted repeat: 8 – **Repeat**, **Consensus**, Mismatch

GCAGGAGCTGTGCCTGGAGAGGCCGACTTGAGGAAGTTTCCGGCACGGAGAGCC
 TGTCGGGGG**GCAGGAGCTGGGCCTGGAGAGGCC**ACCGTGAGGCGTGAGCTGGGC
 CTAATGAGATTAGTGTGAG**GCAGGATCTGGGCCTGTTGAAGCT**GCTGGGAGGCAG
 GCAGGACCTTGTTTGGGGAAGGCCTCTGTGAG**GCAAAAGCTGGGCCTGGAGAGGC**
CATTGTCAC**GCAGGAGCTGGGCCTGTCCAGGCC**ATTGGGAGGCAGAAGGTGGGCC
 TGGAAAGCTTGACTTGAGGCAGTTTGGGGCCTACAAAGGCCAGTAGGAGCTGG**GCA**
GGAGCTGAGCCAAAAGAGGTIGTTCTGAGGCAGGAGTTGGACCTCCAGACACAGT
 CAGAAGAAAGAGCACAGCCTGTAGAGGCCACCTGGGG**GCAGCAGCTGGGCCTAGA**
GATGCCAACGAAAG**GCAGGAGCTGGGCCTTGAGAGGCT**

>chr7_20 | NC_000007_20 | Start: 1396756 | Stop: 1396850 | Evidence Level: 1 | Num. of predicted repeat: 2 – **Repeat**, **Consensus**

TGCTATGGGCCCTTCCAGAGATTCCCCACCCAGCTCCCCTCCAATGGCCAATGTCC
 ATGTCATGGTT**TGCTATGGGCCCTTCCAGAGATTCCCCA**

- **CRISPRCasFinder Predictions:**

>chr13_392 | NC_000013_392 | Start: 109318667 | Stop: 109318921 | Evidence Level: 4 | Num.

of predicted repeat: 5 – Repeat, Degenerate Repeat, Consensus, Mismatch

TTACAGAGAGCTGATTGGCCCATTTTACAGAGAGCTGATTGGTCCATTTTACAGA
GAGCTGATTGGTCCATTTTGACAGGTGCTGATTGGTGAGTTTACAATCCCCGAGTT
AAACACAGAGTGCTGATTGGTGCAAGTTACAATCCTCTAGCTAGACATAACAGTTCT
CCAAGTCCCCACCAGATTAGCTAGATACAGAGTGCTGATTGGTGCAAGCCACAAACC
CCAAGCTAGACACAGAGTGCTGATTGGTGCAAT

>chrX_23 | NC_000023_23 | Start: 3170226 | Stop: 3170464 | Evidence Level: 3 | Num. of

predicted repeat: 5 – Repeat, Degenerate Repeat, Consensus, Mismatch

CACATCCTGCTGATTGGTCCATTTTACAGAGCACTGATTGGTCCATTTTACAGAG
TGCTGATTGGTCTGTTTACAGAGTGCTAATTGGTCTGTTTACAGAGTGCTTAT
TGGTGCGTTTCCAATCCTTTAGCTAGACACAAAAGTTCTGCAAGTCCCCACCCAATT
AGCTAGA CACAGAGTGCTGATTGGTGCGTTTACAAACCTTTAGCTAGA CACAGAG
TGCTGATTGGTGCAATTT

- CRISPRDetect Predictions:

>chr18_255 | NC_000018.10 | Start: 78220037 | Stop: 78220212 | Score: 5.33 | Num. of predicted

repeat: 3 – Repeat A, Repeat B, Consensus, Mismatch

GTTCTCCTTTGCTCTGTAGAGTCCTATTCTCTACCTCATCAGGTAGGTCTGTGT
GTGTGAGAATGTTTGTCTCCTTTGCCCTGTAGAGTCCTATTCTCTACCTTGTC
GTGTAGGTTTGTGTGTGGGCGAGTATTGTCTCCTTTGCCCTGTAGAGTCCTAT
TCCTCTACCT

>chr19_147 | NC_000019.10 | Start: 28434544 | Stop: 28434889 | Score: 4.59 | Num. of predicted

repeat: 6 – Repeat A, Repeat B, Consensus, Mismatch

AGGGGTTCTTGCCCTCTCCCCCAGAAAAGCGGGACTTGCCCTCTAAGGGTGAAGG
 AGAAGGGGTTGAGAGGTTCTTGCCCCCTCTCCCAGAAAAGCGGGACTTGCTGCT
 GACGGTGAAGGACAAGGGTTTGAGGGGTTCTTGCCCCCTCCCCCAGAAAAGTGA
 GACTTGCTGCTAACAGTGAAGGACAAGATGTTGAGGGGTTCTTGCCCCCTCCCC
 AGAAAAGTGGGACTTGCCATTAAAGGGTGAAGGAGAAGGGGTTGAGGGGTTCTT
 GCCCTCCCCCAGAAAAGCAGAGAAGAGGTAGAGACATGGAGAGAAGGGGTGG
 GGTCTTGCCCCCTCCCCCAGAAAAG

- CRISPRCasTyper Predictions:

> chr14_2_1 | NC_000014.9 | Start: 30570622 | Stop: 30570874 | Probability: 0.999 | Num. of

predicted repeat: 4 – Repeat A, Repeat B, Consensus

GTGTGTATATATAGCACATATATATAACACACATATATGTATATATATACATATATAACA
 TATATGGTGTGTATATATGTGTGTATATATAGCACATATATATGGTGTGTGTATATAT
 AGCACATATATATATGGTGTGTGTATATATATATGGTGTGTGTATATATAGCACATATA
 TAT

>chr2_2_1 | NC_000002.12 | Start: 47310846 | Stop: 47311074 | Probability: 0.997 | Num. of

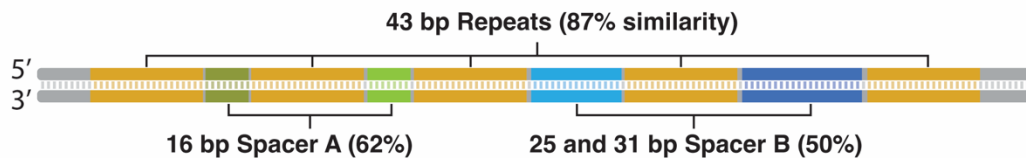
predicted repeat: 4 – Repeat A, Repeat B, Consensus, Mismatch

ATATACACACACATATATATTAGAGAGAGAAATATTGTGAGTGTGGTGTGTGTG
 TGTGAGATATATTTACACACATACACACATATATATAGAGAGAAATATTATATGTG
 CATGTGATGTGTGTCTGAGATATATTTACACACATATATATATAGAGAGAGAAATATT

ATGTGTGTGGTGTGTGTGTGTGAGAGATATTTATATACACACACATATATATATATAT

C

A. *Vicia faba* mitochondrial genome:



B. Mimivirus (APMV) MIMIVIRE loci:

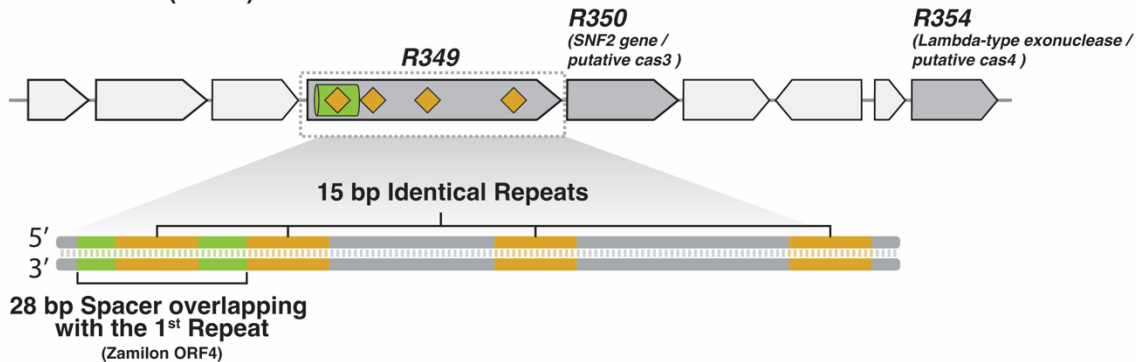


Figure S5.1. Previously reported CRISPR-like elements.

Previously reported CRISPR-like elements in the mitochondrial genome of a plant (*Vicia faba*) and the genome of a giant mimivirus (APMV). A) Array architecture of CRISPR-like sequence from *V. faba* mitochondrion consisted of five similar but non-identical repeats of 43 bp that were interspaced by four small spacers, one each of 31 and 25 bp and two of 16 bp. The two 16 bp spacer sequences were identical for 62% and the two longer spacer sequences were identical for 50%. B) MIMIVIRE loci contains four 15 bp repeats with uneven spacing, located within the coding sequence of uncharacterized R349 ORF of the Mimivirus APMV. The 28 bp matching to the ORF4 gene in the Zamilon genome overlaps with the first repeat of the loci, rather than the spacer sequences. MIMIVIRE lacks Cas1 homolog and encodes only distant, generic homologs of Cas3 and Cas4 annotated as SNF2-containing domain and Lambda-type endonuclease, respectively. Neither the mitochondrial genome nor the genome of a giant mimivirus contain predicted CRISPR loci due to lack of CRISPR-like repeats and array architecture.

CHAPTER SIX

PHYLOGENETIC AND FUNCTIONAL ANALYSIS OF PROKARYOTIC ARGONAUTE
PROTEINS WITH DIVERSE FUNCTIONS

Contribution of Authors and Co-Authors

Manuscript in Chapter 6

Author: Murat Buyukyoruk

Contributions: Designed the research, performed the research, prepared manuscript figures, wrote the manuscript.

Co-Author: Trevor Zahl

Contributions: Designed the research, performed the research.

Co-Author: Blake Wiedenheft

Contributions: Designed the research, wrote the manuscript.

Manuscript Information

Murat Buyukyoruk, Trevor Zahl, Blake Wiedenheft

Status of Manuscript:

X Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

PHYLOGENETIC AND FUNCTIONAL ANALYSIS OF PROKARYOTIC ARGONAUTE
PROTEINS WITH DIVERSE FUNCTIONS

Murat Buyukyoruk¹, Trevor Zahl¹, Blake Wiedenheft¹

¹Department of Microbiology and Immunology, Montana State University, Bozeman, MT
59717, USA

Abstract

Argonaute (Ago) proteins are found in all three domains of life. In Eukaryotes, Argonautes are the catalytic engine of RNA interference (RNAi). Argonautes are also present in prokaryotes and show far greater phylogenetic diversity compared to eukaryotic Argonautes (eAgos). Prokaryotic Argonautes are broken up into three distinct clades: long-A, long-B, and short pAgos. Most of the pAgos that have been experimentally characterized are long-A pAgos, while short pAgos have been understudied. Short pAgos are structurally unique as they are composed of 2 of the 4 domains that long pAgos and eAgos possess. However, previous studies have identified a gene encoded nearby short pAgos that were hypothesized to be “analog of PAZ” (APAZ) genes. Many APAZ genes also contain unique fusions to enzymatically diverse ‘effector’ domains. This diversity in short pAgos highlights a potential sensor/effector mechanism, where the short pAgo acts as sensor that activates enzymatically diverse APAZ-Effectors. However, previous phylogenetic analyses were not comprehensive enough to capture the catalog the extent of APAZ-effector diversity. Here, we demonstrated the intricate domain architecture of 2,167 Agos and 779 APAZ genes revealing previously unappreciated domains that are associated with the prokaryotic argonautes and their APAZ genes. Additionally, we demonstrated that APAZ genes miss a PAZ domain fold while containing the N-terminal domains of a canonical Ago. This expansion provides a complete picture of the diversity of Argonautes and establishes a foundation for exploring their functional implications, potentially revolutionizing our understanding and applications in gene regulation and synthetic biology.

Introduction

Argonaute proteins (Agos) are an ancient family of proteins that are conserved in all domains of life. In eukaryotes, argonautes are best known for their ability to use small RNA guides to recognize and cleave the complementary RNA targets. This process is referred to as RNA interference (RNAi) and is involved in regulating gene expression and protecting against parasitic genetic elements. Previous bioinformatic analyses of prokaryotic genomes revealed that Agos are present in both bacteria and archaea (17% and 25%, respectively)^{1, 2} and that are significantly more diverse in domain architecture than their eukaryotic counterparts³⁻⁵. Additionally, pAgos can use DNA or RNA guides to target RNA or DNA, which is in contrast to eAgos which are strictly RNA-guided RNA-targeting systems. Despite this diversity and widespread conservation of Agos in prokaryotes, few pAgos have been functionally or biochemically characterized to date⁶⁻¹¹.

Prokaryotic argonautes are broken up into three distinct clades based on the domains they possess: long-A, long-B, and short pAgos^{1, 2, 12}. Similar to eAgos, long-A pAgos contain an RNase H-like endonucleolytic PIWI domain with a conserved DEDX (X denotes D, H, or K) tetrad, a MID and PAZ domain responsible for anchoring the 5' and 3' ends of guides, respectively, an N domain involved in strand dissociation after target cleavage, and two linker domains (L1 and L2)². Long-B pAgos have the same domain architecture as Long-A Agos but have mutations in their PIWI domains that abolish their nuclease activities¹³. Short pAgos possess only the MID and PIWI domains and, similar to long-B, have catalytically inactive PIWI domains². However, many pAgos with inactive PIWI domains are adjoined to or located in the same operon as genes containing putative nucleases^{1, 12} that are hypothesized to play roles in either the generation of nucleic acid guides or in the cleavage of complementary targets. While published studies proposed that

eliminating foreign genetic elements through their nucleolytic cleavage may be the primary mode of action for pAgos, it has been hypothesized that these proteins might also be implicated in the regulation of other genetic processes, not necessarily requiring DNA cleavage^{14, 15}.

Most pAgos that have been studied are predominantly categorized within the “long” clades, while short pAgos have been understudied. Short pAgos are structurally unique as they lack the typical PAZ and N domains. However, many short pAgos are flanked by genes that encode a surrogate APAZ domain (analog of PAZ)^{2, 16, 17}. Although the APAZ domain might not exhibit sequence-level homology with the conventional PAZ domains, it's suspected to mimic its structural attributes¹⁶. Notably, as reported by Ryazansky et al., APAZ domains are often fused to enzymatically diverse ‘effector’ domains². These fusions are phylogenetically diverse and are hypothesized to provide unique functionality to their associated pAgos. The most common APAZ fusions include: SIR2 domains¹⁸, TIR domains¹⁹, and DUF4365 domains²⁰ which are homologous to the RecB-like domain of the Mrr subfamily of PD-(D/E)XK nucleases^{2, 17}. These diverse domain fusions highlight the potential versatility of short pAgos in prokaryotic systems. However, past phylogenetic analyses were not comprehensive enough to identify the full suite of APAZ-fused effectors. Here, we use both sequence specific models and structural prediction to comprehensively identify the domain architecture of 2,167 Agos. We also identify 779 APAZ genes with diverse fusions. The unveiling of these effector domains not only enriches our understanding of the versatile functionalities embedded within pAgos but also underscores the adaptive evolution of these proteins in response to varied environmental challenges. Recognizing these diverse domains is imperative as they potentially play pivotal roles in cellular mechanisms, defense strategies, and gene regulatory processes in prokaryotes.

Results and Discussion

Expansion of Ago-associated Effector Domains

To understand the diversity of Agos and their associated effector domains we sourced 25,502 complete bacterial and 398 archaeal complete genomes from the NCBI Assembly database and identified potential Argonaute proteins within protein sequences that containing PIWI and MID domains. Using refined filtering techniques, a non-redundant list of 1,129 eukaryotic and 1,038 prokaryotic argonaute proteins was generated. In parallel, we identified 779 APAZ genes within genomes containing argonautes. Both sets underwent phylogenetic tree construction to discern their evolutionary relationships (**Figure 6.1A**).

Furthermore, we identified non-canonical fusion domains by querying all identified APAZ and Agos with a suite of HMMs. Using this technique, we identified eAgo associated Gly-rich and GAGE domains, long-B pAgo associated OB-fold, HTH (helix-turn-helix) and Alba2 domains, and short pAgo/APAZ associated SIR2, TIR (Toll-interleukin like receptor), DUF4365 domains, all of which have been previously identified (**Figure 6.1B**)^{2, 16, 18-21}. Unsurprisingly, this technique was unable to identify a significant portion of Ago fusions, especially the N-terminal extensions of eAgos and APAZ genes. To address this pitfall, we predicted the structure of Ago and APAZ proteins through AlphaFold²², extracted the fusion domains and used FoldSeek²³ to annotate fused domains. This method was able to identify additional CD-NTase (cGAS/DncV-like nucleotidyltransferase), HATPase (Histidine-kinase like ATPase), HNH endonuclease and Mrr-like SMEK domains that were previously missed through HMM-based domain annotation methods (**Figure 6.1B**). Interestingly, this technique also identified SIR2- and TIR-fused APAZs that contain additional fusions to ATPases and SMEK domains, respectively, suggesting that these

fusions might have different functions compared to the short pAgos containing SIR2 or TIR domains alone. While CD-NTase and HATPase fusions suggest a regulatory function of short pAgos, the HNH endonuclease and Mrr-like SMEK domain fusions imply an alternative way of RNA or DNA targeting mechanisms through nuclease activity of effectors in the presence of inactive PIWI domains in short pAgos.

CD-NTase (Cyclic di-nucleotide forming nucleotidyltransferase) domains are involved in the synthesis of cyclic di-nucleotides, which are secondary messengers in bacterial signaling pathways²⁴. The fusion of CD-NTase domains to APAZ genes could represent an evolutionary strategy in prokaryotes, potentially allowing the APAZ-associated Argonaute proteins to contribute to cellular signaling and regulatory networks. Such a fusion could hypothetically endow the Argonaute-APAZ complex with the ability to detect specific nucleic acid sequences and, in response, synthesize cyclic di-nucleotides that act as intracellular signals. These signals could then modulate gene expression, activate immune responses, or trigger other defense mechanisms. This suggests a multifaceted role for these proteins, where they not only guide the cell in recognizing invasive genetic elements but also play a part in the broader regulatory response to such threats. This indicates that APAZ-associated Argonautes might act as sensors, linking nucleic acid detection to signal transduction, opens up new perspectives on the complexity of prokaryotic defense systems. It also raises the possibility that these proteins could be harnessed in biotechnology, providing a tool for the controlled activation of cellular pathways in response to designed nucleic acid inputs. This could have far-reaching implications for synthetic biology, where such systems might be used to engineer bacteria with sophisticated responses to

environmental cues or to create novel diagnostic tools that respond to the presence of specific genetic markers.

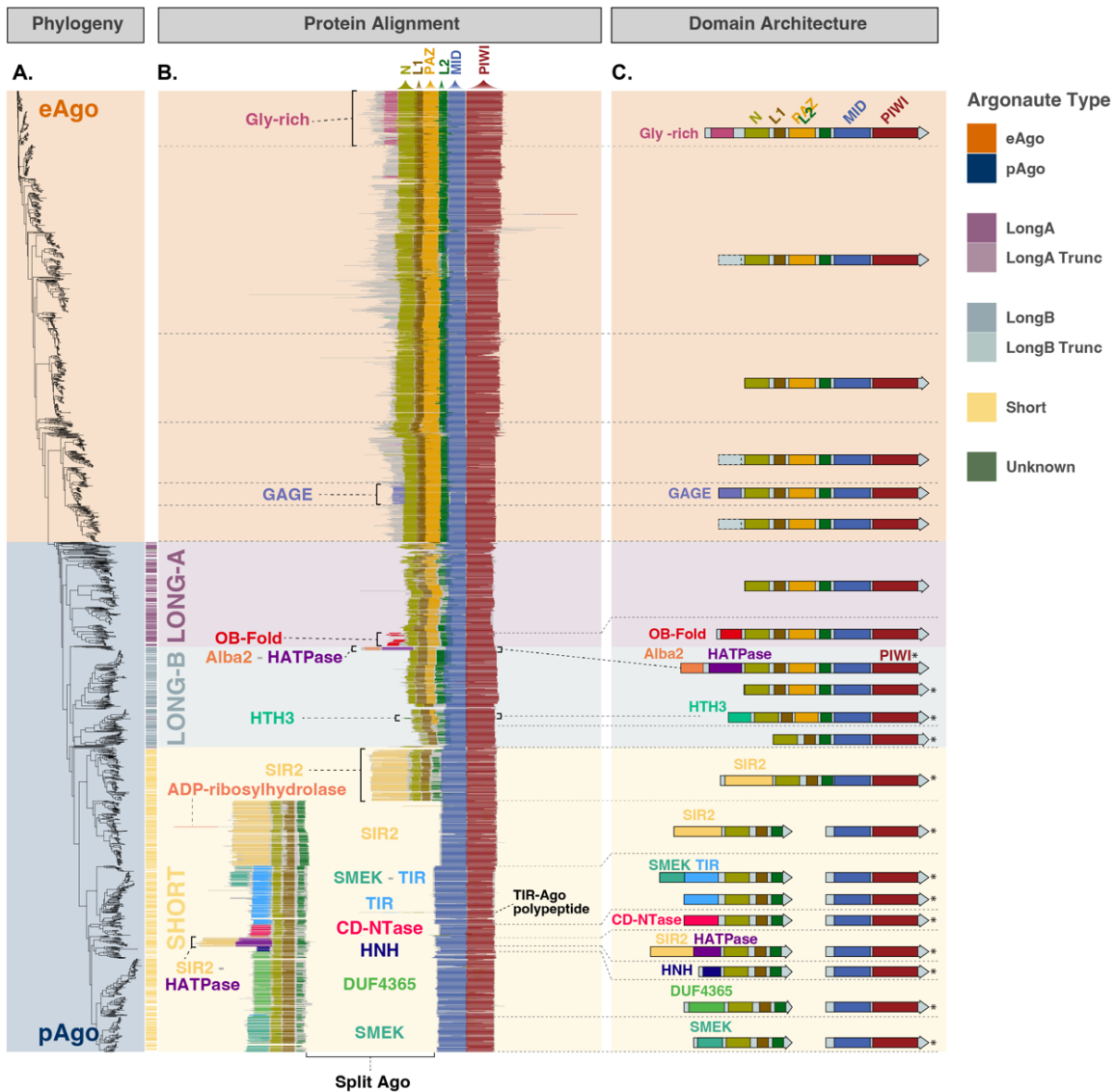


Figure 6.1. Comparative Analysis of Argonaute Protein Architectures.

(A) Phylogenetic representation of Argonaute proteins, categorized into eAgo and pAgo clusters: eAgo, pAgo, LongA, LongB, Short, and Unknown clades depicted to highlight previous nomenclature of Agos (B) Corresponding protein alignment visualizing the conservation across the sequences. Variations in color intensities illustrate amino acid conservation. (C) Domain architecture of the Argonaute proteins, detailing specific domains such as Gly-rich, GAGE, OB-Fold, and HATPase, among others.

Similar to CD-NTase, HATPase domains are also involved in cellular signaling. HATPase (Histidine kinase-like ATPases) represent a class of ATP-binding modules that are pivotal in a variety of cellular processes, serving as engines for energy transduction and integral components of signal transduction pathways²⁵⁻²⁷. These domains are typically found within the cellular machinery that catalyzes ATP hydrolysis, converting chemical energy stored in ATP molecules into mechanical work or signal transduction cascades²⁸. They are integral to the function of various protein families, including histidine kinases, DNA gyrase, Hsp90 chaperones, and MutL, among others²⁹. As such, HATPase domains play a fundamental role in diverse cellular operations, from driving motor proteins and managing cellular stress responses to mediating the repair and replication of DNA. The fusion of an HATPase domain to APAZ genes may point towards a dual functional role for the APAZ-Ago complex, where it acts as a guide for nucleic acid interactions and simultaneously engages in cellular signaling upon target recognition. This fusion implies that the detection of specific genetic sequences could directly initiate signal transduction pathways, perhaps leading to phosphorylation events. Such a system would allow for immediate cellular responses to environmental stimuli or threats, integrating genetic surveillance with the modulation of signal transduction pathways. This multifaceted functionality would equip prokaryotes with the means to not only detect but also react to various cellular states or external pressures, enhancing their survival and adaptive capabilities.

The integration of HNH endonuclease and Mrr-like SMEK domains with APAZ genes in short prokaryotic Argonautes could represent an evolutionary strategy to restore nuclease activity, compensating for the inactive PIWI domains in these proteins. HNH endonucleases are recognized for their nucleic acid cleavage while Mrr-like SMEK domains, part of the PD-(D/E)XK nuclease

family, are implicated in various DNA processing activities^{30, 31}. The hypothesized role of these fused domains may not be to reinstate sequence-specific targeting but rather to invoke an abortive infection mechanism. This cellular defense response could lead to non-specific nucleic acid cleavage, disrupting cellular processes to prevent the spread of infection and protect the microbial community. Such a mechanism would reflect a sophisticated level of immune response, allowing prokaryotes to combat a wide range of genetic threats through a sacrificial yet effective approach.

N-terminal of eAgos Usually Contains Unstructured Regions with NLS and NES Signals

In contrast to the pAgos, we were unable to provide additional annotations the N-terminal extensions of eAgos due to their disordered nature with the exception of previously identified Gly-rich and GAGE domains. Previous studies highlighted that eAgos from *Arabidopsis thaliana* (Ago1) consists of an unstructured N-terminal extensions prior to the N-domain that contains the amino acid sequences that are essentials for the nuclear localization signals (NLS) and nuclear export signals (NES) that facilitate the transportation of these argonaute proteins through the nuclear pore³². To investigate whether these unusual extensions in eAgos contain any localization signals, we extracted the amino acid sequences of eAgo N-terminal extensions and searched for NLS and NES signatures^{33, 34}. From 1,129 sequences searched, we identified 315 NLS and 277 NES sequences that highlights the variation in the eAgo N-terminals is a result of localization instead of unusual domain fusions contrary to the prokaryotic counterparts.

The presence of nuclear localization signals (NLS) and nuclear export signals (NES) in eukaryotic Argonautes (eAgos) is indicative of their dynamic role within the cell. These signals are essential for the intracellular trafficking of eAgos, allowing them to shuttle between the nucleus and cytoplasm. This shuttling is crucial for the diverse functions that eAgos play in gene

regulation. In the nucleus, eAgos are involved in processes such as transcriptional gene silencing, whereas in the cytoplasm, they partake in post-transcriptional gene silencing by interacting with microRNAs and small interfering RNAs.

Unlike prokaryotic Argonautes (pAgos), which often exhibit domain fusions that confer specific enzymatic activities, eAgos seems to rely on their inherent structural flexibility, often characterized by unstructured N-terminal extensions. These extensions allow eAgos to interact with various cellular components, facilitating their import into the nucleus where they can engage in the regulation of chromatin states or process pre-miRNAs. Upon completion of these tasks, NESs direct their export back into the cytoplasm to maintain a constant interplay between the different cellular compartments.

APAZ: Absence of PAZ

The term "APAZ" was originally coined based on the observed proximity of these genes to short pAgos, hinting at a potential analog to the PAZ domain^{2, 16}. Similar to PAZ domains, the APAZ domains were predicted to bind guide nucleic acids and target the mobile genetic elements. While many reports focused on N-terminal domain fusions of APAZ genes (i.e., SIR2, TIR domains), the C-terminal domain of these fusions were remained understudied^{2, 16, 18-21}. In the process of annotating Ago and APAZ fusions, we observed that APAZ genes structurally resembles the N, L1 and L2 domains of long pAgos, while they lack the fold of a PAZ domain entirely (**Figure 6.1B,C**). Furthermore, our analysis revealed that the loss of PAZ domain can be also seen in the pAgos that were previously annotated as long-B pAgos and continues throughout the short pAgo clades, highlighting that APAZ should stand from absence of PAZ instead of analogous of PAZ.

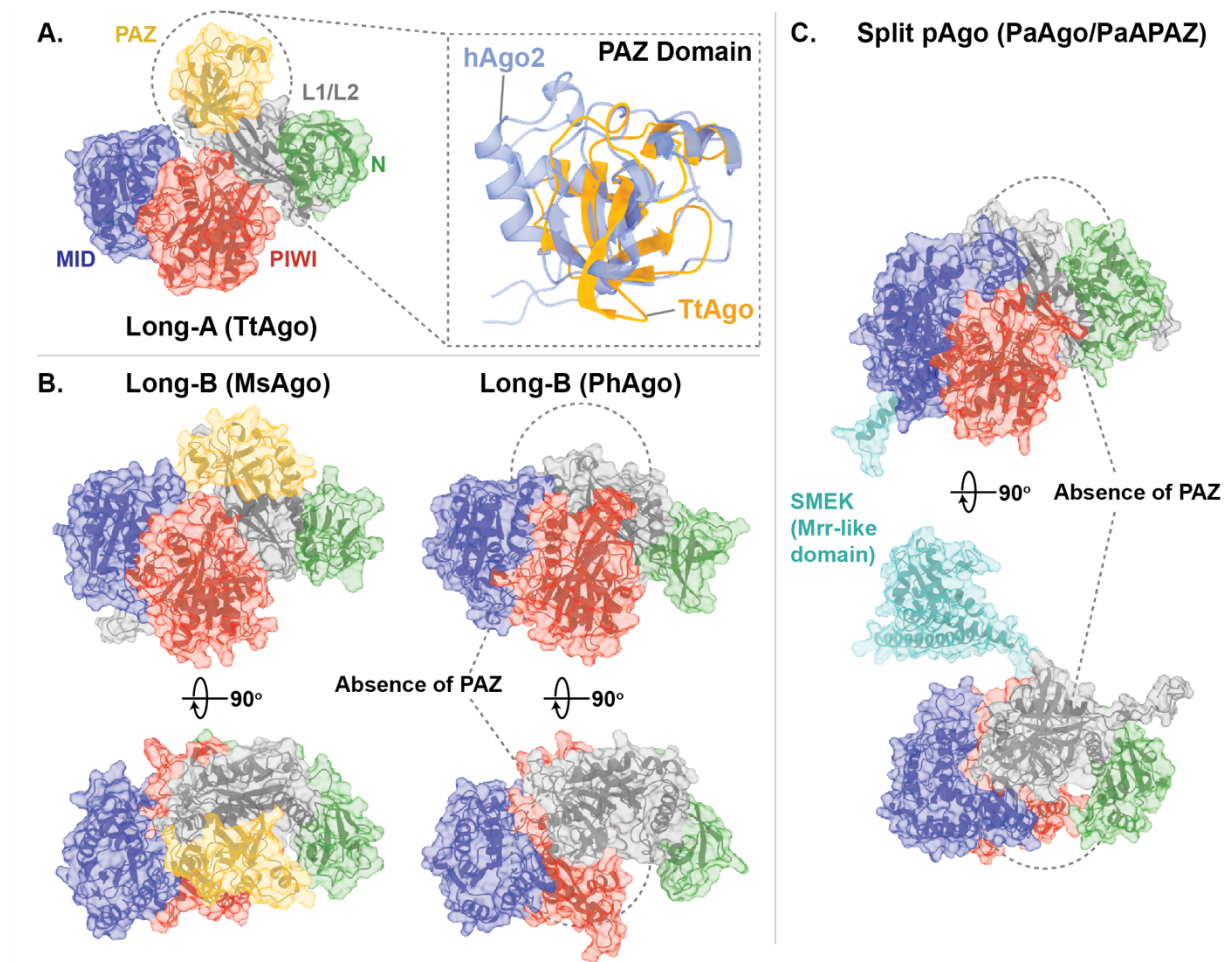


Figure 6.2. Structural variations in prokaryotic argonautes highlighting the presence or absence of the PAZ domain.

(A) Resolved structure of TtAgo with highlighted domains: PIWI in red, MID in blue, PAZ in yellow, N in green and L1/L2 in grey. Superimposition of the PAZ domain of human Ago (hAgo2) with TtAgo from *Thermus thermophilus* of the long-A clade, demonstrating a consistent presence of the PAZ domain. (B) Comparative representation among AlphaFold predictions of long-B pAgos: MsAgo from *Mesoflavibacter* sp. displaying a truncated version and PhAgo from *Pedobacter heparinus* entirely lacking the PAZ domain. (C) Illustration of the defining absence of the PAZ domain in short pAgos, represented by PaAgo from *Pseudomonas aeruginosa*. AlphaFold prediction of PaAgo/APAZ-SMEK heterodimer that can be co-purified together (data not shown). APAZ gene of *Pseudomonas aeruginosa* contains a unique fusion annotated as SMEK that belongs to Mrr-like protein family. The figure emphasizes the evolutionary trajectory of the PAZ domain's presence, starting from its diminished occurrence in the long-B clade to its consistent absence in short pAgos.

To examine the accuracy of our initial observations, we analyzed the predicted structures of prokaryotic argonautes across long-A, long-B, and short pAgos. Specifically, we selected representative structures: TtAgo from *Thermus thermophilus* for the long-A clade, MsAgo from *Mesoflavibacter sp.* and PhAgo from *Pedobacter heparinus* for the long-B clade, and PaAgo from *Pseudomonas aeruginosa* for the short pAgos and its APAZ gene upstream that contains a SMEK domain fusion (**Figure 6.2**). Each proteins sequences of Agos was then submitted to AlphaFold to predict the structure. Superimposing PAZ domain of human Ago (hAgo2) onto these structures revealed that the PAZ domain of hAgo2 aligned well with TtAgo from the long-A clade, highlighting conservation of PAZ domains in long-A pAgos (**Figure 6.2A**). However, among the long-B pAgos, MsAgo had a truncated version of the PAZ domain, while PhAgo completely lacked the PAZ domain (**Figure 6.2B**). Furthermore, the absence of the PAZ domain was a defining characteristic in the short pAgos, as exemplified by PaAgo (**Figure 6.2C**). This suggests an evolutionary trajectory where the absence of the PAZ domain starts in the long-B clade and is consistent in the short pAgos.

Given the role of PAZ domains in binding the 3' end of the guide, long-B and short pAgos lacking the PAZ domain may potentially employ shorter nucleic acid guides. Furthermore, the absence of the PAZ domain might signal a repurposing within these proteins. This shift appears to drive the recruitment of new effector domains like DNA binding domains, SIR2, TIR, and other effectors, not as a direct replacement for the PAZ domain but rather an evolutionary response to its absence. Interestingly, in some instances, these are fused directly as part of the polypeptide, creating a single, continuous protein structure. However, in many short pAgos, these components remain as two distinct subunits, representing a unique 'split pAgo' configuration.

Predicted Modular Mechanisms of Short pAgos in Response to Cellular Stress

Through our phylogenetic analysis we have determined that Argonautes are diverse in form, which leads us to hypothesize that they are diverse in function. Specifically, we hypothesize that short pAgos can utilize multiple unique APAZ genes to respond to unique cellular threats. To test this hypothesis, we queried genomes that contain short pAgos to determine if they encode multiple distinct APAZ genes. Our comprehensive investigation revealed that over 25 genomes possess multiple APAZ genes alongside short pAgos (**Figure S6.1**). Supporting our hypothesis, genomes did not encode redundant APAZ genes, but contained multiple distinct APAZ genes with unique effector fusions like such as APAZ-SIR2 and APAZ-SMEK (**Figure 6.3**). These data suggest that these microbial genomes are equipped to dynamically adapt their defense mechanisms based on diverse stressors and potential invading agents.

Detecting APAZ genes presents its own set of challenges. While identifying APAZ genes adjacent to short pAgos is relatively straightforward, their low sequence similarity makes genome-wide detection more difficult. Relying solely on sequence-specific similarity for identification could result in overlooking a considerable number of these genes. It underscores the potential need for enhanced tools that can perform high-throughput protein predictions based on structure. However, current structural prediction methods demand extensive computational resources, making it prohibitive to predict the structure of every open reading frame (ORF). As research progresses, there is an increasing need for efficient computational tools capable of deciphering the complex genomic landscape of APAZ genes and their associated effectors.

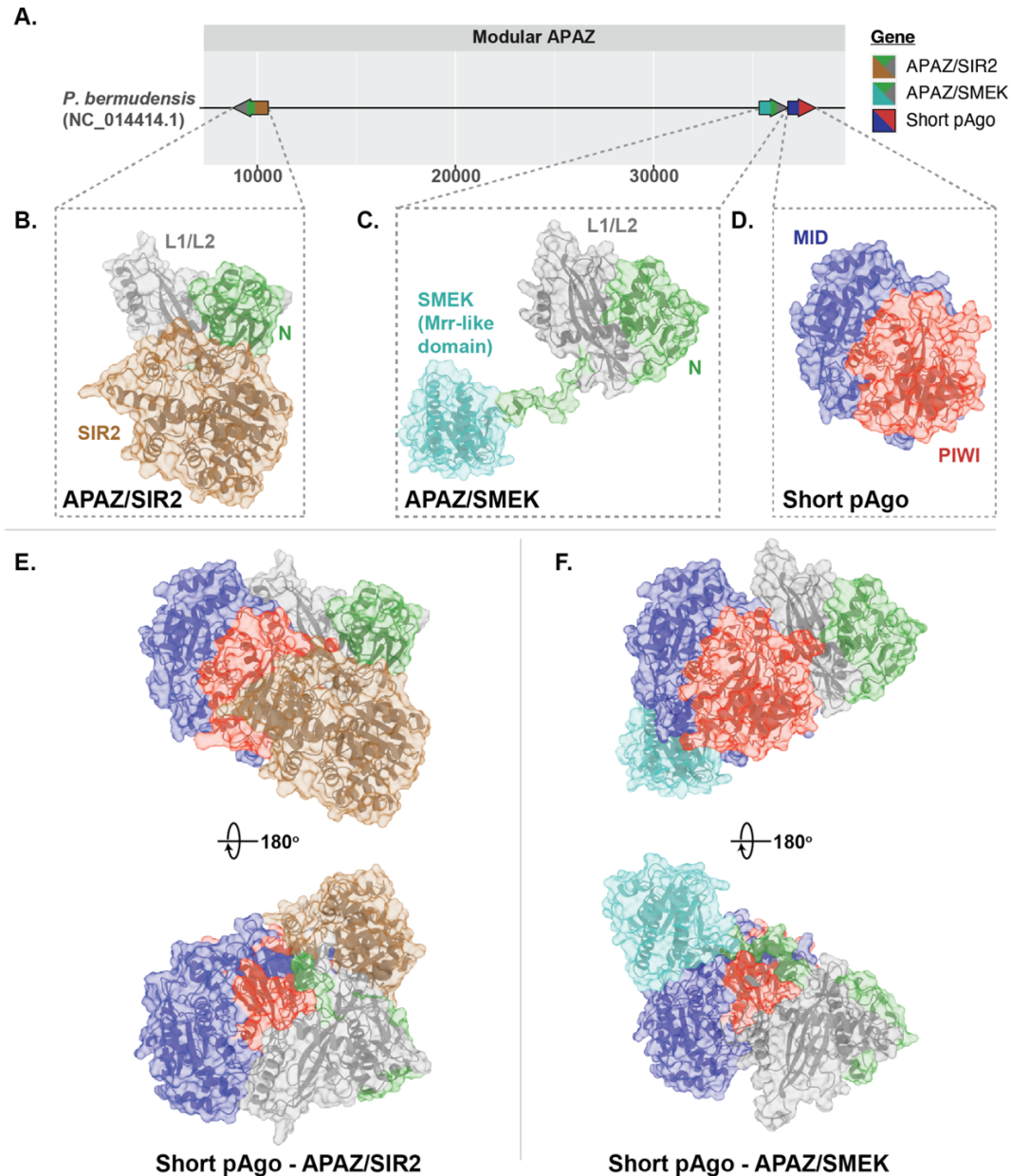


Figure 6.3. Proposed model of modular recruitment of APAZ-effectors by short pAgos.

(A) Gene diagram showcasing the location of APAZ genes and short pAgo from *Parvularcula bermudensis*, with annotated gene regions for APAZ/SIR2, APAZ/SMEK, and Short pAgo. The genomic position is indicated on the x-axis. (B) AlphaFold prediction of APAZ-SIR2 representing the APAZ and SIR2 (in brown). (C) AlphaFold prediction of APAZ-SMEK representing the APAZ in association with SMEK (in cyan). (D) AlphaFold prediction of short pAgo structure representation of the MID (in blue) and PIWI (in red) domains. (E) AlphaFold prediction of short pAgo (depicted with MID in blue and PIWI in red) interacting with both APAZ-SIR2 and (F) APAZ-SMEK.

Conclusion

The work presented here provides a comprehensive analysis of Argonaute proteins, highlighting the intricate diversity of short prokaryotic Argonautes and their associated APAZ domains. The exploration of APAZ domains revealing that these domains frequently lack the typical PAZ fold, suggesting a distinct evolutionary pathway for short pAgos starting from the Argonautes belong to long-B clades. Additionally, our bioinformatics and structural analysis has expanded the known repertoire of Argonaute-associated domains, revealing diverse effector domains that are fused to APAZ genes. We identified CD-NTase, HATPase, HNH endonuclease and Mrr-like SMEK domains in addition to previously reported SIR2, TIR, and DUF4365 domains^{2, 16, 18-21}. The identification of these various APAZ domains fused with various effector domains suggests a complex functional landscape for these proteins, possibly reflecting a modular approach to cellular defense mechanisms in prokaryotes. The study's findings underscore the importance of Argonaute proteins not only in gene regulation but also in the adaptive immune responses of prokaryotes, offering insights into their potential applications in biotechnology. By extending the understanding of Argonaute proteins beyond the well-characterized long clades and into the less understood short pAgos, this research could support new possibilities for utilizing of these proteins in synthetic biology or therapeutic contexts.

Acknowledgement

We thank members of the Wiedenheft laboratory for comments, suggestions, and critical feedback. Research in the Wiedenheft laboratory is supported by National Institutes of Health (R35GM134867), the M.J. Murdock Charitable Trust, a young investigator award from Amgen, a generous gift from the Rosolowsky family, and the Montana State University Agricultural Experimental Station (USDA NIFA)

Methods

Building a Database of Ago Proteins

To generate a list of argonaute proteins, 25,502 complete bacterial and 398 archaeal complete genomes accessed from the NCBI Assembly database on June 10th of 2021. PRODIGAL v2.6.3³⁵ (default parameters) was used to predict all possible ORFs origination from 58,864 chromosome and plasmid sequences of bacterial and archaeal assemblies. In parallel, CDS sequences from a total of 5,152 reference list of eukaryotic genomes accessed from NCBI protein database. To predict the Argonaute proteins in the list of generated prokaryotic ORFs and eukaryotic CDS sequences, we queried all amino acid sequences and searched PIWI and MID domains of argonautes with the HMM models downloaded from PFAM database by using HMMER (hmmsearch) v3.2.1³⁶. A total of 4,428 eAgo and 1,478 pAgo identified that contains both MID and PIWI domains together. A non-redundant list of argonaute proteins generated by CD-HIT v4.8.1³⁷ with a 95% identity cutoff, resulting in a total of 1,129 eukaryotic and 1,038 prokaryotic argonautes.

Identification of APAZ Genes in the Genomes Containing Agos

Previously identified APAZ genes were accessed from the data presented by Ryazansky et al. that contains the list of APAZ genes categorized under five different clades². Each subset of APAZ proteins then aligned with MAFFT v7.429³⁸ with the following command-line options: “*--localpair --maxiterate 1000*”. Resulting alignments then used to generate HMM profiles by using HMMER (hmmbuild) v3.2.1³⁶ with the default parameters. APAZ genes were then searched in the genomes that containing argonaute proteins, resulting with prediction of 910 APAZ genes. A non-redundant list of APAZ proteins generated by CD-HIT v4.8.1³⁷ with a 95% identity cutoff, resulting in a total of 779 APAZ proteins sequences.

Ago and APAZ Phylogenies

A local version of MAFFT v7.429³⁸ was used to align argonaute and APAZ proteins with the following command-line options: “*--localpair --maxiterate 1000*”. Two maximum-likelihood phylogenetic trees for both Agos and APAZs were then generated from aligned sequences, using FastTree 2.1.11³⁹. The trees were generated using the following parameters “*-wag -gamma -pseudo*”. Generated phylogenetic trees then linked based on the accession number with cophyloplot package available on RStudio.

Domain Annotation of Argonaute and APAZ Sequences

In addition to the HMM profiles for MID and PIWI domain, various HMMs were accessed from PFAM database, including N, L1, PAZ, L2, and previously identified APAZ fusion domains to annotate the previously known domains that are associated with Ago and APAZ proteins. A local version of AlphaFold2 v.2.3.0²² was used to predict the structure of 2,167 Ago and 779 APAZ proteins. Resulting predicted structures were then used to extract unannotated regions of each

protein by ChimeraX⁴⁰ and a custom pychimera script⁴¹. Extracted regions queried for function prediction by a local version of Foldseek v.6.29²³ that compares the queried structure to the available PDB database. Identified domains were then illustrated on the phylogenetic tree by using ggtree, ggplot2 and tidyverse^{42, 43}.

References

1. Swarts, D.C., et al., *The evolutionary journey of Argonaute proteins*. Nat Struct Mol Biol, 2014. **21**(9): p. 743-53.
2. Ryazansky, S., A. Kulbachinskiy, and A.A. Aravin, *The Expanded Universe of Prokaryotic Argonaute Proteins*. Mbio, 2018. **9**(6).
3. Wang, Y., et al., *Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex*. Nature, 2008. **456**(7224): p. 921-6.
4. Yuan, Y.R., et al., *Crystal structure of A-aeolicus Argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage*. Molecular Cell, 2005. **19**(3): p. 405-419.
5. Parker, J.S., et al., *Enhancement of the Seed-Target Recognition Step in RNA Silencing by a PIWI/MID Domain Protein*. Molecular Cell, 2009. **33**(2): p. 204-214.
6. Zander, A., et al., *Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii*. Nat Microbiol, 2017. **2**: p. 17034.
7. Swarts, D.C., et al., *Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA*. Nucleic Acids Res, 2015. **43**(10): p. 5120-9.
8. Olovnikov, I., et al., *Bacterial argonaute samples the transcriptome to identify foreign DNA*. Mol Cell, 2013. **51**(5): p. 594-605.
9. Swarts, D.C., et al., *Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute*. Mol Cell, 2017. **65**(6): p. 985-998 e6.
10. Jolly, S.M., et al., *Thermus thermophilus Argonaute Functions in the Completion of DNA Replication*. Cell, 2020. **182**(6): p. 1545-1559 e18.
11. Hegge, J.W., et al., *DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute*. Nucleic Acids Research, 2019. **47**(11): p. 5809-5821.
12. Makarova, K.S., et al., *Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements*. Biol Direct, 2009. **4**: p. 29.
13. Wang, Y., et al., *Structure of the guide-strand-containing argonaute silencing complex*. Nature, 2008. **456**(7219): p. 209-13.
14. Swarts, D.C., *Prokaryotic Argonautes Function beyond Immunity by Unlinking Replicating Chromosomes*. Cell, 2020. **182**(6): p. 1381-1383.

15. Lisitskaya, L., A.A. Aravin, and A. Kulbachinskiy, *DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins*. Nat Commun, 2018. **9**(1): p. 5165.
16. Makarova, K.S., et al., *Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements*. Biology Direct, 2009. **4**.
17. Koopal, B., S.K. Mutte, and D.C. Swarts, *A long look at short prokaryotic Argonautes*. Trends in Cell Biology, 2023. **33**(7): p. 605-618.
18. Zaremba, M., et al., *Short prokaryotic Argonautes provide defence against incoming mobile genetic elements through NAD depletion*. Nature Microbiology, 2022. **7**(11): p. 1857-+.
19. Koopal, B., et al., *Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA*. Cell, 2022. **185**(9): p. 1471-+.
20. Kim, S.Y., Y. Jung, and D. Lim, *Argonaute system of is a heterodimeric nucleic acid-guided nuclease*. Biochemical and Biophysical Research Communications, 2020. **525**(3): p. 755-758.
21. Xu, Y., et al., *The N-terminal extension of Arabidopsis ARGONAUTE 1 is essential for microRNA activities*. Plos Genetics, 2023. **19**(3).
22. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold*. Nature, 2021. **596**(7873): p. 583-589.
23. van Kempen, M., et al., *Fast and accurate protein structure search with Foldseek*. Nat Biotechnol, 2023.
24. Kranzusch, P.J., *cGAS and CD-NTase enzymes: structure, mechanism, and evolution*. Curr Opin Struct Biol, 2019. **59**: p. 178-187.
25. Stewart, R.C., *Protein histidine kinases: assembly of active sites and their regulation in signaling pathways*. Current Opinion in Microbiology, 2010. **13**(2): p. 133-141.
26. Dutta, R., L. Qin, and M. Inouye, *Histidine kinases: diversity of domain organization*. Molecular Microbiology, 1999. **34**(4): p. 633-640.
27. Buschiazzo, A. and F. Trajtenberg, *Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level?* Annual Review of Microbiology, Vol 73, 2019. **73**: p. 507-+.
28. Cai, E.P., et al., *Histidine Kinase SlnI and cAMP/PKA Signaling Pathways Antagonistically Regulate Mating and Virulence via Transcription Factor PrfI*. Journal of Fungi, 2021. **7**(8).

29. Perna, L., et al., *AlphaFold predicted structure of the Hsp90-like domains of the neurodegeneration linked protein sarsin reveals key residues for ATPase activity*. *Frontiers in Molecular Biosciences*, 2023. **9**.
30. Kala, S., et al., *HNH proteins are a widespread component of phage DNA packaging machines*. *Proceedings of the National Academy of Sciences of the United States of America*, 2014. **111**(16): p. 6022-6027.
31. Bujnicki, J.M. and L. Rychlewski, *Identification of a PD-(D/E)XK-like domain with a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs*. *Gene*, 2001. **267**(2): p. 183-191.
32. Xu, Y., et al., *The N-terminal extension of Arabidopsis ARGONAUTE 1 is essential for microRNA activities*. *PLoS Genet*, 2023. **19**(3): p. e1010450.
33. Bernhofer, M., et al., *NLSdb-major update for database of nuclear localization signals and nuclear export signals*. *Nucleic Acids Research*, 2018. **46**(D1): p. D503-D508.
34. Xu, D.R., et al., *LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins*. *Bioinformatics*, 2015. **31**(9): p. 1357-1365.
35. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification*. *BMC Bioinformatics*, 2010. **11**: p. 119.
36. Prakash, A., et al., *The HMMER Web Server for Protein Sequence Similarity Search*. *Curr Protoc Bioinformatics*, 2017. **60**: p. 3 15 1-3 15 23.
37. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*. *Bioinformatics*, 2006. **22**(13): p. 1658-9.
38. Katoh, K. and D.M. Standley, *MAFFT: iterative refinement and additional methods*. *Methods Mol Biol*, 2014. **1079**: p. 131-46.
39. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix*. *Mol Biol Evol*, 2009. **26**(7): p. 1641-50.
40. Meng, E.C., et al., *UCSF ChimeraX: Tools for structure building and analysis*. *Protein Sci*, 2023. **32**(11): p. e4792.
41. Rodriguez-Guerra Pedregal, J. and J.D. Marechal, *PyChimera: use UCSF Chimera modules in any Python 2.7 project*. *Bioinformatics*, 2018. **34**(10): p. 1784-1785.
42. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures*. *Curr Protoc Bioinformatics*, 2020. **69**(1): p. e96.

43. Yu, G., et al., *Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree*. Mol Biol Evol, 2018. **35**(12): p. 3041-3043.

Supplemental Figures

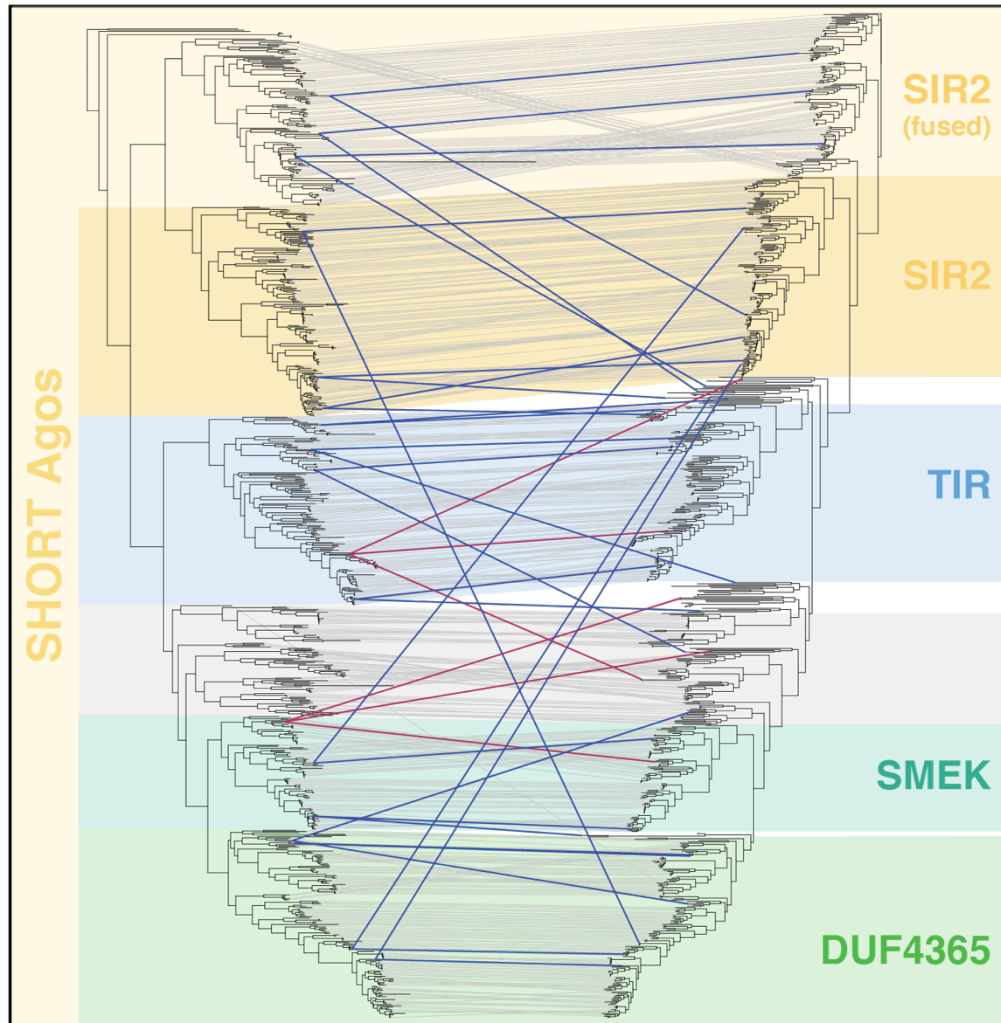


Figure S6.1. Illustration of the relationship between short pAgos and multiple APAZ effectors. Using genome accession numbers, we connected APAZ genes to short Agos. On the left, a phylogeny of short pAgos grouped by their associated APAZ genes observed adjacent to the genes in the same operon. On the right, phylogeny of APAZ genes identified in the genomes of bacterial and archaeal sequences that contains Argonaute proteins. Blue line represents the short pAgos localized with two APAZ-effectors, while red line represents the short pAgos localized with three APAZ-effectors.

CHAPTER SEVEN

GENERAL CONCLUSION AND OUTLOOK

Bacteriophages are the most abundant biological entities and their consistent threat to bacterial populations has instigated an arms race of survival. This has led bacteria to evolve diverse defense mechanisms, and phages to subsequently counteract those. At the forefront of these defense systems stand the CRISPR-Cas systems, evolutionary marvels of adaptive immunity in prokaryotes. These systems have been increasingly recognized, not just for their defense roles but also for their potential in biotechnological applications, particularly in areas like gene editing, diagnostics, and DNA/RNA recording.

From CRISPR Interference to Regulation of CRISPRs

While there's been significant attention directed towards understanding the interference mechanism of CRISPR systems against DNA phages, my initial research trajectory was drawn towards a different narrative: defense against RNA phages. This choice was informed by the contrasting profiles of viruses that invade eukaryotic versus prokaryotic systems. In eukaryotes, RNA viruses are predominant. However, in the realm of bacteriophages, DNA phages take center stage. This divergence poses intriguing questions: Have bacteria and archaea evolved specialized defenses against RNA, which remain largely undiscovered, or are RNA phages simply so scarce in prokaryotic environments that a dedicated RNA-targeted defense system wasn't evolutionarily favored?

To unravel this mystery, my endeavors initially ventured into investigations of the claims around the Type I-F CRISPR system (**Chapter 2**). Although traditionally recognized as a DNA-

targeting defense, some evidence suggested its potential to also target RNA, albeit via a distinct mechanism. While further research is paramount to fully understand the alleged RNA-targeting capabilities of this system in *Pseudomonas aeruginosa*, my attempts to repurpose the model for RNA phage immunity did not yield affirmative results, hinting at uncharted territories in the RNA defense landscape.

Such findings necessitated a pivot in my research direction. I chose to conduct an in-depth study on a significant, yet previously overlooked component of the CRISPR-Cas systems: the CRISPR leader sequences. Many existing CRISPR detection tools focused solely on either the repetitive array sequence or *cas* genes, neglecting the wealth of information contained within the CRISPR leader. By integrating this neglected segment into our analyses, I aimed to unveil a more complete view of CRISPR systems, encompassing their evolutionary diversity and further expanding their potential in biotechnological applications.

During my Ph.D. work, while the overarching role of the CRISPR leader in regulating transcription was recognized, its profound implications for the process of integrating foreign DNAs and its impact on the evolution of the CRISPR loci began to surface only recently. In our exploration, we revealed that a significant number of CRISPR loci had leaders from I-C, I-E, I-F and II-C showcasing sequence motifs identified by the integration host factor (IHF) proteins (**Chapter 3**). Astonishingly, many of these leaders also carried previously unidentified upstream sequence motifs within a conserved spacing, which in turn, influenced the recruitment of these upstream sequence motifs by the Cas1-2 integrase. The importance of the spacing these motifs became evident as we hypothesized and subsequently confirmed its critical role in their presentation to the CRISPR-associated adaptation proteins (**Chapter 3**). Additionally, we

recognized that, while many type I-E CRISPR leaders belong to *Pseudomonadota* incorporated IHF-binding sites, a significant amount of leaders originated from *Actinomycetoma* did not. Instead, they exhibited unique DNA sequence motifs, highlighting an intriguing possibility suggesting a synchronized evolution of these leader sequences with CRISPR-Cas systems and distinct host factors. Most importantly, it highlighted for me that there was an opportunity to expand the CRISPR discovery process itself by including more information from the leaders. Through rigorous categorizations rooted in 16S rRNA-based taxonomy and CRISPR subtyping, we identified 77 distinct motifs across more than 24,000 CRISPR leader belongs to type I, II, III and V (**Chapter 4**). Their significance was revealed when some of these motifs showed alignment with sequences previously associated with key CRISPR functions. This discovery reinforced their regulatory roles and could offer a promising avenue to improve CRISPR detection and classification methodologies and enhance the detection of CRISPRs (**Chapter 4**).

Looking at Genomes Outside of the Prokaryotes to Find Evidence of CRISPRs

A somewhat unexpected turn in my research journey led me to investigate the possible existence of CRISPR loci in the human genome, termed "hCRISPRs" (**Chapter 5**). Using the bioinformatic tools that I developed, I conducted an in-depth analysis of the human reference genome, aiming to rigorously detect and evaluate so-called hCRISPRs, if they existed. This investigation unearthed numerous potential CRISPR arrays that diverged from the canonical CRISPR structures known from prokaryotes. Despite the initial large number of potential arrays, few survived rigorous confidence assessments. Further attempts to correlate these to validated *cas* genes yielded no tangible matches, raising further doubts about the legitimacy of the reported hCRISPRs (**Chapter 5**). This was unsurprising, considering that repetitive DNA sequences can

constitute anywhere from 1-80% of eukaryotic genomes. When I expanded this search to include a broader selection of eukaryotic genomes and even prokaryotes using non-strict search parameters, the results echoed the initial findings. Such outcomes underscore the importance of systematic analysis of detected putative arrays. Employing non-strict settings without a meticulous review can lead to the misidentification of genuine CRISPRs, underscoring the need for careful discernment in distinguishing true CRISPRs from other repetitive sequences (**Chapter 5**).

Diversity of Prokaryotic Argonautes

The latter phase of my research focused on the Argonaute proteins, especially the lesser-explored prokaryotic Argonautes. With advanced bioinformatics and molecular biology techniques as my compass, I embarked on an expedition into the depths of these proteins, unraveling their genetic associations, relationships with APAZ-effectors, and potential defensive mechanisms (**Chapter 6**). We showed that the diversity observed in eAgos are distinct from pAgos. The N-terminal extensions in eAgos are usually unstructured regions that encode NLS or NES motifs, the N-terminal extensions in pAgos and short pAgo associated APAZ genes were encode various DNA-binding or effector domains. In addition to the previously reported SIR2, TIR and DUF4365 domains, we identified effector domains that fused to the APAZ domain, including SMEK domains (a distinct domain belongs to the Mrr-like protein family), CD-NTase domains and HNH fold-encoding domains. Furthermore, we expanded the domain architecture of SIR2 and TIR domain-containing APAZ genes, by showing they can possess various combinations of these domains, such as SMEK/TIR and SIR2/HATPase (**Chapter 6**). Finally, we investigated the possibility that, upon a target recognition, short pAgos might act as a sensors that recruit various APAZ-effector proteins

encoded in the genome, depending on the encountered stress, and thereby serve as a possible modular switching mechanism (**Chapter 6**).

Concluding Remarks

Ultimately, the realm of bacteriophages and bacterial interactions stands not just as a testament to nature's ingenuity but also as a repository of possibilities for human innovation. Each chapter of my research journey has been a step towards understanding this vast realm a little better, and through that understanding, gleaning tools and techniques that hold the promise of revolutionizing biotechnology.

By embracing the often-overlooked nuances of bacterial defense mechanisms, like the CRISPR leader and the prokaryotic Argonautes, I've endeavored to contribute to a more complete understanding of this intricate field. It's my hope that these contributions pave the way for future innovations, firmly rooting them in a comprehensive understanding of nature's most intricate defense systems. In this evolutionary dance between phages and bacteria, as they navigate the challenges and counter-challenges thrown their way, may we continue to find inspiration, solutions, and a deeper appreciation for life.

CUMULATIVE REFERENCES CITED

- Abby, S.S., et al., MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, 2014. 9(10): p. e110726.
- Abudayyeh, O. O. & Gootenberg, J. S. CRISPR diagnostics. *Science* (1979) 372, 914–915 (2021).
- Aeling KA, Opel ML, Steffen NR, Tretyachenko-Ladokhina V, Hatfield GW, Lathrop RH, Senear DF. Indirect Recognition in Sequence-specific DNA Binding by *Escherichia coli* Integration Host Factor. *J Biol Chem*. 2006 Dec;281(51):39236–48.
- Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J.W., Choi, P.M., Kitajima, M., Simpson, S.L., Li, J., et al. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci Total Environ* 728, 138764.
- Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth Phase-Dependent Variation in Protein Composition of the *Escherichia coli* Nucleoid. *J. Bacteriol.* 181, 6361–6370 (1999).
- Alkhnbashi OS, Meier T, Mitrofanov A, et al. CRISPR-Cas bioinformatics. *Methods* 2020;172(3-11, doi:10.1016/j.ymeth.2019.07.013
- Alkhnbashi, O. S. et al. Characterizing leader sequences of CRISPR loci. *Bioinformatics* 32, i576–i585 (2016).
- Allan-Blitz, L. T. & Klausner, J. D. A Real-World Comparison of SARS-CoV-2 Rapid Antigen Testing versus PCR Testing in Florida. *J Clin Microbiol* 59, (2021).
- Alonso, J. C., Weise, F. & Rojo, F. The *Bacillus subtilis* Histone-like Protein Hbsu Is Required for DNA Resolution and DNA Inversion Mediated by the β Recombinase of Plasmid pSM19035. *J. Biol. Chem.* 270, 2938–2945 (1995).
- Athukoralage, J. S. et al. The dynamic interplay of host and viral enzymes in type iii crispr-mediated cyclic nucleotide signalling. *Elife* 9, (2020).
- Athukoralage, J. S., Graham, S., Grüşchow, S., Rouillon, C. & White, M. F. A Type III CRISPR Ancillary Ribonuclease Degrades Its Cyclic Oligoadenylate Activator. *J Mol Biol* 431, 2894–2899 (2019).
- Athukoralage, J. S., Rouillon, C., Graham, S., Grüşchow, S. & White, M. F. Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. *Nature* 2018 562:7726 562, 277–280 (2018).

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009 Jul 1;37(Web Server):W202–8.

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings Int Conf Intell Syst Mol Biol.* 1994;2:28–36.

Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015;6(11, doi:10.1186/s13100-015-0041-9

Barrangou, R. et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* 315, 1709–1712 (2007).

Bedford, T., Greninger, A.L., Roychoudhury, P., Starita, L.M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., et al. (2020). Cryptic transmission of SARS-CoV-2 in Washington State. medRxiv.

Béguin, P., Chekli, Y., Sezonov, G., Forterre, P. & Krupovic, M. Sequence motifs recognized by the casposon integrase of *Aciduliprofundum boonei*. *Nucleic Acids Res.* 47, 6386–6395 (2019).

Benz, F. et al. Type IV-A3 CRISPR-Cas systems drive inter-plasmid conflicts by acquiring spacers in trans. <http://biorxiv.org/lookup/doi/10.1101/2023.06.23.546257> (2023) doi:10.1101/2023.06.23.546257.

Bernheim A, Bikard D, Touchon M, Rocha EPC. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic Acids Res.* 2019 Nov 20;48(2):748–60.

Biggerstaff, M., Jhung, M.A., Reed, C., Fry, A.M., Balluz, L., and Finelli, L. (2014). Influenza-like illness, the time to seek healthcare, and influenza antiviral receipt during the 2010-2011 influenza season-United States. *J Infect Dis* 210, 535-544.

Biswas A, Staals RH, Morales SE, et al. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* 2016;17(356, doi:10.1186/s12864-016-2627-0

Bolotin A, Quinkis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology.* 2005 Aug 1;151(8):2551–61.

Bonilla N, Rojas MI, Netto Flores Cruz G, Hung SH, Rohwer F, Barr JJ. Phage on tap-a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ.* 2016;4:e2261.

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science.* 2008;321(5891):960-4. 263

- Brown, M. W. et al. Assembly and translocation of a CRISPR-Cas primed acquisition complex. *bioRxiv* 41, 1–11 (2017).
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (Oxford, England). 2009;25(15):1972-3.
- Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, Glover CVC, Graveley BR, Terns RM, Terns MP. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *S. treptococcus thermophilus*. *Mol Microbiol*. 2014 Jul;93(1):98–112.
- Castillo, A.E., Parra, B., Tapia, P., Acevedo, A., Lagos, J., Andrade, W., Arata, L., Leal, G., Barra, G., Tambley, C., et al. (2020). Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J Med Virol*.
- Cavazzana-Calvo, M. et al. Gene Therapy of Human Severe Combined Immunodeficiency (SCID)-X1 Disease. *Science* 288, 669–672 (2000).
- Cech, T.R. and J.A. Steitz, The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 2014. 157(1): p. 77-94.
- Chalmers, R., Guhathakurta, A., Benjamin, H. & Kleckner, N. IHF Modulation of Tn10 Transposition: Sensory Transduction of Supercoiling Status via a Proposed Protein/DNA Molecular Spring. *Cell* 93, 897–908 (1998).
- Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* 360, 436 (2018).
- Chiang, H.R., et al., Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*, 2010. 24(10): p. 992-1009.
- Chowdhury, S., et al., Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell*, 2017. 169(1): p. 47-57 e11.
- Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25(11):1422-1423, doi:10.1093/bioinformatics/btp163
- Connolly M, Arra A, Zvoda V, Steinbach PJ, Rice PA, Ansari A. Static Kinks or Flexible Hinges: Multiple Conformations of Bent DNA Bound to Integration Host Factor Revealed by Fluorescence Lifetime Measurements. *J Phys Chem B*. 2018 Dec 13;122(49):11519–34.

Consortium CeS. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998;282(5396):2012-8, doi:10.1126/science.282.5396.2012

Couvin, D. et al. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* 46, W246–W251 (2018).

Crawley, A. B., Henriksen, J. R. & Barrangou, R. CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems. *The CRISPR Journal* 1, 171–181 (2018).

Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004 Jun;14(6):1188–90.

Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–90 (2004).

Czech, B. and G.J. Hannon, Small RNA sorting: matchmaking for Argonautes. *Nat Rev Genet*, 2011. 12(1): p. 19-31.

Dai, X., et al., In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature*, 2017. 541(7635): p. 112-116.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165 (2011).

Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* 3, 945 (2012).

Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O.G., Faria, N.R., Wang, C., Yu, G., Bushnell, B., Pan, C.-Y., et al. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 369, 582-587.

Deveau, H. et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390–1400 (2008).

Dhingra, Y., Suresh, S. K., Juneja, P. & Sashital, D. G. PAM binding ensures orientational integration during Cas4-Cas1-Cas2-mediated CRISPR adaptation. *Mol. Cell* 82, 4353-4367.e6 (2022).

Drain, P. K. Rapid Diagnostic Testing for SARS-CoV-2. <https://doi.org/10.1056/NEJMcp2117115> (2022) doi:10.1056/NEJMCP2117115.

East-Seletsky, A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 2016 538:7624 538, 270–273 (2016).

- Eckweiler D, Dudek C-A, Hartlich J, Brötje D, Jahn D. PRODORIC2: the bacterial gene regulation database in 2018. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D320–6.
- Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113.
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D* 66, 486–501 (2010).
- Fagerlund, R. D. et al. Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *Proc. Natl. Acad. Sci.* 114, 201618421 (2017).
- Fauver, J.R., Petrone, M.E., Hodcroft, E.B., Shioda, K., Ehrlich, H.Y., Watts, A.G., Vogels, C.B.F., Brito, A.F., Alpert, T., Muyombwe, A., et al. (2020). Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* 181, 990-996 e995.
- Feher T, Karcagi I, Blattner FR, Posfai G. Bacteriophage recombineering in the lytic state using the lambda red recombinases. *Microbial biotechnology.* 2012;5(4):466-76.
- Fineran, P. C. et al. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 111, (2014).
- Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39, W29–W37 (2011).
- Fire, A., et al., Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 1998. 391(6669): p. 806-11.
- Flamand MC, Goblet JP, Duc G, et al. Sequence and Transcription Analysis of Mitochondrial Plasmids Isolated from Cytoplasmic Male-Sterile Lines of *Vicia-Faba*. *Plant Mol Biol* 1992;19(6):913-923.
- Friedman, R.C., et al., Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 2009. 19(1): p. 92-105.
- Fromm, B., et al., A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet*, 2015. 49: p. 213-42.
- Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012 Dec;28(23):3150–2.
- Fu L, Xie C, Jin Z, Tu Z, Han L, Jin M, et al. The prokaryotic Argonaute proteins enhance homology sequence-directed recombination in bacteria. *Nucleic Acids Res.* 2019;47(7):3568-79.
- Garcia-Doval, C. et al. Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6. *Nature Communications* 2020 11:1 11, 1–9 (2020).

Garg, S., Kim, L., Whitaker, M., O'Halloran, A., Cummings, C., Holstein, R., Prill, M., Chai, S.J., Kirley, P.D., Alden, N.B., et al. (2020). Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. *MMWR Morb Mortal Wkly Rep* 69, 458-464.

Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* 27, 14–25 (2018).

Goemann CLC, Wilkinson R, Henriques W, et al. Genome sequence, phylogenetic analysis, and structure-based annotation reveal metabolic potential of *Chlorella* sp. SLA-04. *Algal Research* 2023;69.

Gong, B., et al., Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A*, 2014. 111(46): p. 16359-64.

Gonzalez-Reiche, A.S., Hernandez, M.M., Sullivan, M.J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J., Khan, Z., et al. (2020). Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*.

Gootenberg, J. S. et al. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a and Csm6. *Science* (1979) 360, 439–444 (2018).

Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* (1979) 356, 438–442 (2017).

Goren, M. G. et al. Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. *Cell Rep.* 16, 2811–2818 (2016).

Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: Maximizing usable data in an alignment. *BMC Bioinformatics* 8, 1–8 (2007).

Grant, C.E., T.L. Bailey, and W.S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 2011. 27(7): p. 1017-8.

Grubaugh, N.D., Gangavarapu, K., Quick, J., Matteson, N.L., De Jesus, J.G., Main, B.J., Tan, A.L., Paul, L.M., Brackney, D.E., Grewal, S., et al. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology* 20.

Grüschow, S., Grüschow, G., Adamson, C. S. & White, M. F. Specificity and sensitivity of an RNA targeting type III CRISPR complex coupled with a NucC endonuclease effector. *Nucleic Acids Res* (2021) doi:10.1093/NAR/GKAB1190.

Gudbergdottir S, Deng L, Chen Z, Jensen JVK, Jensen LR, She Q, Garrett RA. Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol Microbiol.* 2011 Jan;79(1):35–49.

- Guo, T.W., et al., Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell*, 2017. 171(2): p. 414-426 e12.
- Hale, C. R. et al. RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* 139, 945–956 (2009).
- Haley, B. and P.D. Zamore, Kinetic analysis of the RNAi enzyme complex. *Nat Struct Mol Biol*, 2004. 11(7): p. 599-606.
- Halpin-Healy TS, Klompe SE, Sternberg SH, Fernández IS. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature*. 2020 Jan 18 [cited 2020 Jan 28];577(7789):271–4.
- Haniford, D. B. Transpososome Dynamics and Regulation in Tn10 Transposition. *Crit. Rev. Biochem. Mol. Biol.* 41, 407–424 (2006).
- Hart, O.E., and Halden, R.U. (2020). Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. *Science of The Total Environment* 730.
- Hegge JW, Swarts DC, Chandradoss SD, Cui TJ, Kneppers J, Jinek M, et al. DNAGuided DNA cleavage at moderate temperatures by *Clostridium butyricum* Argonaute. *Nucleic Acids Res.* 2019;47(11):5809-21.
- Hershey DA, ed. *The Bacteriophage Lambda*. Cold Spring Harbor Laboratory Press. 1971.
- Herzik, M. A., Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* 10, 1–9 (2019).
- Heussler GE, Miller JL, Price CE, Collins AJ, O'Toole GA. Requirements for *Pseudomonas aeruginosa* type I-F CRISPR-Cas adaptation determined using a biofilm enrichment assay. *J Bacteriol.* 2016;198(22):3080–90.
- Hickman, A. B. & Dyda, F. Mechanisms of DNA transposition. *Mob. DNA III* 529–553 (2015) doi:10.1128/9781555819217.ch25.
- Hindson, J. (2020). COVID-19: faecal-oral transmission? *Nat Rev Gastroenterol Hepatol*.
- Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., et al. (2020). First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine* 382, 929-936.
- Hoyland-Kroghsbo, N.M., K.A. Munoz, and B.L. Bassler, Temperature, by Controlling Growth Rate, Regulates CRISPR-Cas Activity in *Pseudomonas aeruginosa*. *mBio*, 2018. 9(6).

Hrle, A., et al., Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biol*, 2013. 10(11): p. 1670-8.

<https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/in-vitro-diagnostics-euas>.

Hu, C. et al. Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* 598, 515–520 (2021).

Hubley R, Finn RD, Clements J, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res* 2016;44(D1):D81-9, doi:10.1093/nar/gkv1272

Hufford MB, Seetharam AS, Woodhouse MR, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 2021;373(6555):655-662, doi:doi:10.1126/science.abg5289

Huo, Y. et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* 21, 771–777 (2014).

Huyke, D. A. et al. Enzyme Kinetics and Detector Sensitivity Determine Limits of Detection of Amplification-Free CRISPR-Cas12 and CRISPR-Cas13 Diagnostics. *Anal Chem* 94, 9826–9834 (2022).

Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119 (2010).

Ishino Y, Krupovic M, Forterre P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J Bacteriol* 2018;200(7), doi:10.1128/JB.00580-17

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology* 169, 5429–5433 (1987).

Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R., and Kuroda, M. (2020). A proposal of alternative primers for the ARTIC Network’s multiplex PCR to improve coverage of SARS-CoV-2 genome sequencing. *bioRxiv*.

Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJJ. CRISPR-Cas: Adapting to change. *Science* (80-). 2017 Apr 7;356(6333):eaal5056.

Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol.* 1961 Jun;3(3):318–56.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36, 338-345.

Jansen R, van Embden JDA, Gaastra W, et al. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 2002;43(6):1565-1575, doi:DOI 10.1046/j.1365-2958.2002.02839.x

Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, et al. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol.* 2009;394(4):644-52.

Jessica L. Prince-Guerra, P. O. A. et al. Evaluation of Abbott BinaxNOW Rapid Antigen Test for SARS-CoV-2 Infection at Two Community-Based Testing Sites — Pima County, Arizona, November 3–17, 2020. <https://www.cdc.gov/mmwr/volumes/70/wr/pdfs/mm7003e3-H.pdf>.

Jia, N., Jones, R., Yang, G., Ouerfelli, O. & Patel, D. J. CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA^{>p} Formation Terminating RNase Activity. *Mol Cell* 75, 944-956.e6 (2019).

Jiao, C. et al. Noncanonical crRNAs derived from host transcripts enable multiplexable RNA detection by Cas9. *Science* (1979) 372, 941–948 (2021).

Jinek, M., et al., A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 2012. 337(6096): p. 816-21.

Jonas, S. and E. Izaurralde, Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet*, 2015. 16(7): p. 421-33.

Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc.* 2012;7(8):1511-22.

Kamashev D, Agapova Y, Rastorguev S, Talyzina AA, Boyko KM, Korzhenevskiy DA, Vlaskina A, Vasilov R, Timofeev VI, Rakitina T V. Comparison of histone-like HU protein DNA-binding properties and HU/IHF protein sequence alignment. *PLoS One.* 2017;12(11):1–24.

Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nature Biomedical Engineering* 2021 5:7 5, 643–656 (2021).

Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005;33(2):511-8.

Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol.* 2013 Apr 1;30(4):772–80.

Kaya E, Doxzen KW, Knoll KR, Wilson RC, Strutt SC, Kranzusch PJ, et al. A bacterial Argonaute with noncanonical guide RNA specificity. *Proceedings of the National Academy of Sciences*. 2016;113(15):4057-62.

Kazlauskienė, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* (1979) 357, 605–609 (2017).

Kieper SN, Almendros C, Brouns SJJ. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiol Lett*. 2019 Jun 1;366(11):2016–20.

Kim JG, Garrett S, Wei Y, Graveley BR, Terns MP. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR–Cas system. *Nucleic Acids Res*. 2019 Sep 19;47(16):8632–48.

Kim S, Loeff L, Colombo S, Jergic S, Brouns SJJ, Joo C. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature*. 2020;579(7797):141–5.

Kim, V.N., J. Han, and M.C. Siomi, Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol*, 2009. 10(2): p. 126-39.

Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature*. 2019 Jul 12 [cited 2019 Aug 6];571(7764):219–25.

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* 571, 219–225 (2019).

Kobryn K, Naigamwalla DZ, Chaconas G. Site-specific DNA binding and bending by the *Borrelia burgdorferi* Hbb protein. *Mol Microbiol*. 2000 Jul;37(1):145–55.

Koonin EV. Evolution of RNA- and DNA-guided antiviral defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biology Direct*. 2017;12(1):5.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*.

Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*. 8, R61 (2007).

Künne, T. et al. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol. Cell* 63, 852–864 (2016).

Kupczok A, Landan G, Dagan T. The Contribution of Genetic Recombination to CRISPR Array Evolution. *Genome Biol Evol*. 2015;7(7):1925–39.

- Kuzmenko A, Yudin D, Ryazansky S, Kulbachinskiy A, Aravin AA. Programmable DNA cleavage by Ago nucleases from mesophilic bacteria *Clostridium butyricum* and *Limnithrix rosea*. *Nucleic Acids Res.* 2019;47(11):5822-36.
- La Rosa, G., Iaconelli, M., Mancini, P., Bonanno Ferraro, G., Veneri, C., Bonadonna, L., Lucentini, L., and Suffredini, E. (2020). First detection of SARS-CoV-2 in untreated wastewaters in Italy. *Sci Total Environ* 736, 139652.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.
- Larremore, D. B. et al. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv* 7, (2021).
- Lau, R. K. et al. Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity. *Mol Cell* 77, 723-733.e6 (2020).
- Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., and Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Ann Intern Med* 172, 577-582.
- Lavoie, B. D. & Chaconas, G. Site-specific HU binding in the Mu transpososome: conversion of a sequence-independent DNA-binding protein into a chemical nuclease. *Genes Dev.* 7, 2510–2519 (1993).
- Laxmikanthan, G. et al. Structure of a holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. *eLife* 5, 1–23 (2016).
- Lee DH, Schleif RF. In vivo DNA loops in *araCBAD*: size limits and helical repeat. *Proc Natl Acad Sci.* 1989 Jan 1;86(2):476–80.
- Lee KZ, Mechikoff MA, Kikla A, Liu A, Pandolfi P, Gimble FS, et al. NgAgoenhanced homologous recombination in *E. coli* is mediated by DNA endonuclease activity. *bioRxiv.* 2019:597237.
- Lee SH, Turchiano G, Ata H, Newsheer S, Romito M, Lou Z, et al. Failure to detect DNA-guided genome editing using *Natronobacterium gregoryi* Argonaute. *Nat Biotechnol.* 2016;35(1):17-8.
- Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. *Trends Biochem. Sci.* 1–13 (2022) doi:10.1016/j.tibs.2022.02.004.
- Levasseur A, Bekliz M, Chabriere E, et al. MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. *Nature* 2016;531(7593):249-52, doi:10.1038/nature17146

- Levine, E. and T. Hwa, Small RNAs establish gene expression thresholds. *Curr Opin Microbiol*, 2008. 11(6): p. 574-9.
- Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet*. 2014 Jul 10;15(7):453–68.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*. 2006;22(13):1658-9.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094-3100.
- Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res*. 42, 2483–2492 (2014).
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S.M., Lau, E.H.Y., Wong, J.Y., et al. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* 382, 1199-1207.
- Li, R., et al., Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Res*, 2016. 26(12): p. 1273-1287.
- Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659 (2006).
- Liao, C. et al. Spacer prioritization in CRISPR–Cas9 immunity is enabled by the leader RNA. *Nat. Microbiol.* (2022) doi:10.1038/s41564-022-01074-3.
- Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. Sect. Struct. Biol.* 75, 861–877 (2019).
- Lillestøl, R. K. et al. CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties: Archaeal CRISPRs of *Sulfolobus*. *Molecular Microbiology* 72, 259–272 (2009).
- Linheiro, R. S. & Bergman, C. M. Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster* P-element. *Nucleic Acids Res*. 36, 6199–6208 (2008).
- Lintner, N.G., et al., Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem*, 2011. 286(24): p. 21643-56.
- Lisitskaya L, Aravin AA, Kulbachinskiy A. DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nat Commun*. 2018;9(1):5165.

Little, A. J., Corbett, E., Ortega, F. & Schatz, D. G. Cooperative recruitment of HMGB1 during V(D)J recombination through interactions with RAG1 and DNA. *Nucleic Acids Res.* 41, 3289–3301 (2013).

Liu CJ, Jiang H, Wu L, Zhu LY, Meng E, Zhang DY. OEPR Cloning: an Efficient and Seamless Cloning Strategy for Large- and Multi-Fragments. *Sci Rep.* 2017;7:44648.

Liu, D., Haniford, D. B. & Chalmers, R. M. H-NS mediates the dissociation of a refractory protein-DNA complex during Tn10/IS10 transposition. *Nucleic Acids Res.* 39, 6660–6668 (2011).

Liu, L. et al. The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* 170, 714-726.e10 (2017).

Liu, T. Y. et al. Accelerated RNA detection using tandem CRISPR nucleases. *Nature Chemical Biology* 2021 17:9 17, 982–988 (2021).

Lobocka MB, Rose DJ, Plunkett G, 3rd, Rusin M, Samojedny A, Lehnher H, et al. Genome of bacteriophage P1. *J Bacteriol.* 2004;186(21):7032-68.

Lodder, W., and de Roda Husman, A.M. (2020). SARS-CoV-2 in wastewater: potential health risk, but also data source. *Lancet Gastroenterol Hepatol.*

Lopez-Alvarez, D., Parra, B., and Cuellar, W.J. (2020). Genome Sequence of SARS-CoV-2 Isolate Cali-01, from Colombia, Obtained Using Oxford Nanopore MinION Sequencing. *Microbiology Resource Announcements* 9, e00573-00520.

Luria SE, Adams JN, Ting RC. Transduction of lactose-utilizing ability among strains of *E. coli* and *S. dysenteriae* and the properties of the transducing phage particles. *Virology.* 1960;12:348-90.

Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ. Structural basis for 5'-endspecific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature.* 2005;434(7033):666-70.

Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 468, 326–329 (2010).

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat Rev Microbiol.* 2015 Nov 28;13(11):722–36.

Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research.* 2013;41(8):4360-77.

Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology Direct.* 2009;4(1):29.

Makarova, K. S. et al. Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antiviral defense. *Nucleic Acids Res* 48, 8828–8847 (2020).

Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* 18, 67–83 (2020).

Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? *CRISPR J.* 1, 325–336 (2018).

Mallapaty, S. (2020). How sewage could reveal true scale of coronavirus outbreak. *Nature*.

Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 2004;32(Web Server issue):W327-31.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 2011;39(Database issue):D225-9.

Marraffini, L.A., CRISPR-Cas immunity in prokaryotes. *Nature*, 2015. 526(7571): p. 55-61.

Masse, E., F.E. Escorcia, and S. Gottesman, Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. *Genes Dev*, 2003. 17(19): p. 2374-83.

Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* 152, 36–51 (2005).

Matsumoto N, Nishimasu H, Sakakibara K, Nishida KM, Hirano T, Ishitani R, et al. Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. *Cell*. 2016;167(2):484-97.e9. 262

Matthews BJ, Dudchenko O, Kingan SB, et al. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* 2018;563(7732):501-507, doi:10.1038/s41586-018-0692-z

McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* 36, 344–355 (1950).

McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol Cell*. 2016 Nov;64(3):616–23.

McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol. Cell* 64, 616–623 (2016).

McMahon, S. A. et al. Structure and mechanism of a Type III CRISPR defence DNA nuclease activated by cyclic oligoadenylate. *Nature Communications* 2020 11:1 11, 1–11 (2020).

- Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., and Brouwer, A. (2020). Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environmental Science & Technology Letters* 7, 511-516.
- Medvedeva, S. et al. Virus-borne mini-CRISPR arrays are involved in interviral conflicts. *Nature Communications* 10, 1–10 (2019).
- Meister, G., Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet*, 2013. 14(7): p. 447-59.
- Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37, 1530–1534 (2020).
- Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682 (2022).
- Mohanraju P, Makarova KS, Zetsche B, et al. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* 2016;353(6299):aad5147, doi:10.1126/science.aad5147
- Mojica FJ, Díez-Villasenor C, Soria E, et al. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* 2000;36(1):244-6, doi:10.1046/j.1365-2958.2000.01838.x
- Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*. 2005 Feb;60(2):174–82.
- Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155, 733–740 (2009).
- Mojica, F. J. M., Díez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* 36, 244–246 (2000).
- Molina, R. et al. Structure of Csx1-cOA4 complex reveals the basis of RNA decay in Type III-B CRISPR-Cas. *Nature Communications* 2019 10:1 10, 1–14 (2019).
- Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The Mu transpososome structure sheds light on DDE recombinase evolution. *Nature* 491, 413–417 (2012).
- Mouw KW, Rice PA. Shaping the *Borrelia burgdorferi* genome: crystal structure and binding properties of the DNA-bending protein Hbb. *Mol Microbiol*. 2007 Mar;63(5):1319–30.

- Musharova, O. et al. Prespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. *Proc. Natl. Acad. Sci.* 118, e2021291118 (2021).
- Nasef, M. et al. Regulation of cyclic oligoadenylate synthesis by the *Staphylococcus epidermidis* Cas10-Csm complex. *RNA* 25, 948–962 (2019).
- Nash, H. A. & Robertson, C. A. Purification and properties of the *Escherichia coli* protein factor required for lambda integrative recombination. *J. Biol. Chem.* 256, 9246–9253 (1981).
- Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments*. *Bioinformatics*, 2009. 25(10): p. 1335-7.
- Nethery, M. A. et al. CRISPRclassify: Repeat-Based Classification of CRISPR Loci. *CRISPR J.* 4, 558–574 (2021).
- Nethery, M. A. et al. CRISPRclassify: Repeat-Based Classification of CRISPR Loci. *The CRISPR Journal* 4, 558–574 (2021).
- Nicholls, R. A. Conformation-independent comparison of protein structures. (2011).
- Nicholson, T. J. et al. Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. *RNA Biol.* 16, 566–576 (2019).
- Nickel, L. et al. Two CRISPR-Cas systems in *Methanosarcina mazei* strain Göl display common processing features despite belonging to different types I and III. *RNA Biology* 10, 779–791 (2013).
- Niewoehner, O. et al. Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* 2017 548:7669 548, 543–548 (2017).
- Nivala J, Shipman SL, Church GM. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat Microbiol.* 2018 Mar 29;3(3):310–8.
- Norais, C., Moisan, A., Gaspin, C. & Clouet-d’Orval, B. Diversity of CRISPR systems in the euryarchaeal Pyrococcales. *RNA Biology* 10, 659–670 (2013).
- Núñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell.* 2016 Jun;62(6):824–33.
- Núñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* 62, 824–833 (2016).
- Núñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* 527, 535–538 (2015).

- Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science* 2022;376(6588):44-53, doi:10.1126/science.abj6987
- Nussenzweig, P. M. & Marraffini, L. A. Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria. *Annu. Rev. Genet.* 54, 93–120 (2020).
- Olovnikov I, Chan K, Sachidanandam R, Newman Dianne K, Aravin Alexei A. Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA. *Molecular Cell.* 2013;51(5):594-605.
- Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings.
- Pardee, K. et al. Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* 165, 1255–1266 (2016).
- Parker, J.S., et al., Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. *Mol Cell*, 2009. 33(2): p. 204-14.
- Peccia, J., Zulli, A., Brackney, D.E., Grubaugh, N.D., Kaplan, E.H., Casanovas-Massana, A., Ko, A.I., Malik, A.A., Wang, D., Wang, M., et al. (2020). SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. *medRxiv*.
- Pérez-Martín J, Rojo F, de Lorenzo V. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol Rev.* 1994;58(2):268–90.
- Petassi MT, Hsieh S-C, Peters JE. Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons. *Cell.* 2020 Dec;183(7):1757-1771.e18.
- Peters JE, Makarova KS, Shmakov S, Koonin E V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci.* 2017 Aug 29;114(35):E7358–66.
- Petit, M.-A., Ehrlich, D. & Jannière, L. pAM β 1 resolvase has an atypical recombination site and requires a histone-like protein HU. *Mol. Microbiol.* 18, 271–282 (1995).
- Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 2021;30(1):70–82.
- Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 30, 70–82 (2021).
- Phok, K. et al. Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon *Pyrococcus abyssi*. *BMC Genomics* 12, 312 (2011).

- Pinilla-redondo, R., Mayo-muñoz, D., Russel, J. & Garrett, R. A. Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. (2019).
- Pougach, K. et al. Transcription, processing and function of CRISPR cassettes in *Escherichia coli*: Transcription and function of *E. coli* CRISPR locus. *Molecular Microbiology* 77, 1367–1379 (2010).
- Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. 2005 Mar 1;151(3):653–63.
- Pourcel C, Touchon M, Villeriot N, et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* 2020;48(D1):D535-D544, doi:10.1093/nar/gkz915
- Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* 151, 653–663 (2005).
- Prakash A, Jeffryes M, Bateman A, et al. The HMMER Web Server for Protein Sequence Similarity Search. *Curr Protoc Bioinformatics* 2017;60(3 15 1-3 15 23, doi:10.1002/cpbi.40
- Prasad, D.V.V., S. Madhusudanan, and S. Jaganathan. uCLUST - A NEW ALGORITHM FOR CLUSTERING UNSTRUCTURED DATA. 2015.
- Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. *PLoS One*. 2010 Mar 10;5(3):e9490.
- Price, M.N., P.S. Dehal, and A.P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*, 2009. 26(7): p. 1641-50.
- Przybilski R, Richter C, Gristwood T, Clulow JS, Vercoe RB, Fineran PC. Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biol*. 2011 May 27;8(3):517–28.
- Pucker B, Holtgrawe D, Stadermann KB, et al. A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. *PLoS One* 2019;14(5):e0216233, doi:10.1371/journal.pone.0216233
- Pul Ü, Wurm R, Arslan Z, Geißen R, Hofmann N, Wagner R. Identification and characterization of *E. coli* CRISPR- cas promoters and their silencing by H-NS. *Mol Microbiol*. 2010 Mar;75(6):1495–512.
- Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14, 290–296 (2017).

Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* 17, 1214–1221 (2020).

Qiu D, Damron FH, Mima T, Schweizer HP, Yu HD. PBAD-Based Shuttle Vectors for Functional Analysis of Toxic and Highly Regulated Genes in *Pseudomonas* and *Burkholderia* spp. and Other Bacteria. *Appl Environ Microbiol.* 2008 Dec 1;74(23):7422–6.

Quick, J. (2020). nCoV-2019 sequencing protocol. protocolsio.

Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols* 12, 1261-1276.

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228-232.

Ramachandran, A., Summerville, L., Learn, B. A., DeBell, L. & Bailey, S. Processing and integration of functionally oriented prespacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. *J. Biol. Chem.* 295, 3403–3414 (2020).

Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simon, P., Allende, A., and Sanchez, G. (2020). SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Res* 181, 115942.

Razvi, A. & Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein Science* 15, 1569–1578 (2006).

Rice PA, Yang S, Mizuuchi K, Nash HA. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell.* 1996 Dec;87(7):1295–306.

Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell* 87, 1295–1306 (1996).

Richter C, Dy RL, McKenzie RE, Watson BNJ, Taylor C, Chang JT, McNeil MB, Staals RHJ, Fineran PC. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* 2014;42(13):8516–26.

Richter C, Gristwood T, Clulow JS, Fineran PC. In Vivo Protein Interactions and Complex Formation in the *Pectobacterium atrosepticum* Subtype I-F CRISPR/Cas System. *PLoS One.* 2012;7(12).

Richter, C. et al. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* 42, 8516–8526 (2014).

Rimoldi, S.G., Stefani, F., Gigantiello, A., Polesello, S., Comandatore, F., Mileto, D., Maresca, M., Longobardi, C., Mancon, A., Romeri, F., et al. (2020). Presence and vitality of SARS-CoV-2 virus in wastewaters and rivers. medRxiv.

Rohs, R. et al. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* 79, 233–269 (2010).

Rojo, F. & Alonso, J. C. The β recombinase of plasmid pSM19035 binds to two adjacent sites, making different contacts at each of them. *Nucleic Acids Res.* 23, 3181–3188 (1995).

Rollie C, Graham S, Rouillon C, White MF. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res.* 2018 Feb 16;46(3):1007–20.

Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res.* 46, 1007–1020 (2018).

Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Research* 46, 1007–1020 (2018).

Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. & White, M. F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife* 4, 1–19 (2015).

Rollins MF, Chowdhury S, Carter J, Golden SM, Wilkinson RA, Bondy-Denomy J, Lander GC, Wiedenheft B. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci.* 2017 Apr 24 [cited 2017 Dec 9];114(26):201616395.

Rollins, M. F. et al. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc. Natl. Acad. Sci.* 114, 201616395 (2017).

Rollins, M. F. et al. Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol. Cell* 74, 132-142.e5 (2019).

Rollins, M.F., et al., Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci U S A*, 2017. 114(26): p. E5113-E5121.

Rollins, M.F., et al., Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from *Pseudomonas aeruginosa*. *Nucleic Acids Res.* 2015. 43(4): p. 2216-22.

Rollins, M.F., et al., Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell*, 2019. 74(1): p. 132-142 e5.

Rostøl, J. T. & Marraffini, L. A. Non-specific degradation of transcripts promotes plasmid clearance during type III-A CRISPR–Cas immunity. *Nature Microbiology* 2019 4:4 4, 656–662 (2019).

Rostøl, J. T. et al. The Card1 nuclease provides defence during type III CRISPR immunity. *Nature* 2021 590:7847 590, 624–629 (2021).

Rouillon, C., Athukoralage, J. S., Graham, S., Grüşchow, S. & White, M. F. Control of cyclic oligoadenylate synthesis in a type III CRISPR system. *Elife* 7, (2018).

Rowland, S.-J., Stark, W. M. & Boocock, M. R. Sin recombinase from *Staphylococcus aureus*: synaptic complex architecture and transposon targeting: Sin recombinase. *Mol. Microbiol.* 44, 607–619 (2002).

Russel J, Pinilla-Redondo R, Mayo-Munoz D, et al. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J* 2020;3(6):462-469, doi:10.1089/crispr.2020.0059

Ryazansky S, Kulbachinskiy A, Aravin AA. The Expanded Universe of Prokaryotic Argonaute Proteins. *mBio*. 2018;9(6):e01935-18.

Sagendorf, J. M., Markarian, N., Berman, H. M. & Rohs, R. DNAProDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.* gkz889 (2019) doi:10.1093/nar/gkz889.

Santiago-Frangos, A. et al. CRISPR-Cas, Argonaute proteins and the emerging landscape of amplification-free diagnostics. *Methods* 205, 1–10 (2022).

Santiago-Frangos, A. et al. Intrinsic signal amplification by type III CRISPR-Cas systems provides a sequence-specific SARS-CoV-2 diagnostic. *Cell Rep Med* 2, 100319 (2021).

Santiago-Frangos, A. et al. Protein-mediated genome folding allosterically enhances site-specific integration of foreign DNA into CRISPRs. <http://biorxiv.org/lookup/doi/10.1101/2023.05.26.542337> (2023) doi:10.1101/2023.05.26.542337.

Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Curr. Biol.* 1–10 (2021) doi:10.1016/j.cub.2021.05.068.

Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Current Biology* 1–10 (2021) doi:10.1016/j.cub.2021.05.068.

Santiago, J. G. Inconsistent treatments of the kinetics of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) impair assessment of its diagnostic potential. *QRB Discov* 3, e9 (2022).

Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* 9, 853–854 (2012).

Schleif R. DNA Looping. *Annu Rev Biochem.* 1992 Jun 8;61(1):199–223.

Schmidt F, Cherepkova MY, Platt RJ. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature*. 2018 Oct 3 [cited 2018 Oct 30];562(7727):380–5.

Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 562, 380–385 (2018).

Schneider VA, Graves-Lindsay T, Howe K, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;27(5):849-864, doi:10.1101/gr.213611.116

Schneider, T. D. & Stephens, R. M. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100 (1990).

Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494, 489–491 (2013).

Semenova, E. et al. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. *Proc. Natl. Acad. Sci.* 113, 7626–7631 (2016).

Sherrill-Mix, S., taxonomizr: Functions to Work with NCBI Accessions and Taxonomy. 2023.

Shipman SL, Nivala J, Macklis JD, Church GM. Molecular recordings by directed CRISPR spacer acquisition. *Science* (80-). 2016 Jul 29;353(6298):aaf1175.

Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* 353, aaf1175 (2016).

Shmakov, S. A. et al. CRISPR Arrays Away from cas Genes. *CRISPR Journal* 3, 535–549 (2020).

Silas, S., et al., Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*, 2016. 351(6276): p. aad4234.

Silverman, J.D., Hupert, N., and Washburne, A.D. (2020). Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. *Sci Transl Med*.

Sinkunas, T. et al. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* 30, 1335–1342 (2011).

Smalakyte, D. et al. Type III-A CRISPR-associated protein Csm6 degrades cyclic hexa-adenylate activator using both CARF and HEPN domains. *Nucleic Acids Res* 48, 9204–9217 (2020).

Sridhara, S., Goswami, H. N., Whyms, C., Dennis, J. H. & Li, H. Virus detection via programmable Type III-A CRISPR-Cas systems. doi:10.1038/s41467-021-25977-7.

Steens, J. A. et al. SCOPE enables type III CRISPR-Cas diagnostics using flexible targeting and stringent CARF ribonuclease activation. *Nature Communications* 2021 12:1 12, 1–12 (2021).

Stefanelli, P., Faggioni, G., Lo Presti, A., Fiore, S., Marchi, A., Benedetti, E., Fabiani, C., Anselmo, A., Ciammaruconi, A., Fortunato, A., et al. (2020). Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Euro Surveill* 25.

Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* 365, 48–53 (2019).

Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF. Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J Mol Biol.* 2009;394(4):653-80.

Suloway, C. et al. Automated molecular microscopy: The new Legimon system. *J. Struct. Biol.* 151, 41–60 (2005).

Sunghyeok Y, Taegeun B, Kyoungmi K, Omer H, Seung Hwan L, Yoon Young K, et al. DNA-dependent RNA cleavage by the *Natronobacterium gregoryi* Argonaute. *bioRxiv*. 2017.

Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932 (2015).

Swarts DC, Hegge JW, Hinojo I, Shiimori M, Ellis MA, Dumrongkulraksa J, et al. Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Research*. 2015;43(10):5120-9. 261

Swarts DC, Jore MM, Westra ER, Zhu Y, Janssen JH, Snijders AP, et al. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*. 2014;507(7491):258-61.

Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, et al. The evolutionary journey of Argonaute proteins. *Nature structural & molecular biology*. 2014;21(9):743-53.

Swarts DC, Szczepaniak M, Sheng G, Chandradoss SD, Zhu Y, Timmers EM, et al. Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. *Mol Cell*. 2017;65(6):985-98.e6.

Swinger KK, Rice PA. IHF and HU: flexible architects of bent DNA. *Curr Opin Struct Biol*. 2004 Feb;14(1):28–35.

Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* 14, 793–796 (2017).

- Tang, A., Tong, Z.D., Wang, H.L., Dai, Y.X., Li, K.F., Liu, J.N., Wu, W.J., Yuan, C., Yu, M.L., Li, P., et al. (2020). Detection of Novel Coronavirus by RT-PCR in Stool Specimen from Asymptomatic Child, China. *Emerg Infect Dis* 26.
- Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;13(1):36-46, doi:10.1038/nrg3117
- Udriste AA, Iordachescu M, Ciceoi R, et al. Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the Ve Locus. *Int J Mol Sci* 2022;23(17), doi:10.3390/ijms23179750
- Vale PF, Little TJ. CRISPR-mediated phage resistance and the ghost of coevolution past. *Proc R Soc B Biol Sci.* 2010 Jul 22;277(1691):2097–103.
- van Dijk M, Bonvin AMJJ. 3D-DART: A DNA structure modelling server. *Nucleic Acids Res.* 2009;37(SUPPL. 2):235–9.
- van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 83, 104351.
- van Gent, D. C., Hiom, K., Paull, T. T. & Gellert, M. Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J.* 16, 2665–2670 (1997).
- van Riet J, Saha C, Strepis N, et al. CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue. *Commun Biol* 2022;5(1):338, doi:10.1038/s42003-022-03249-4
- Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291(5507):1304-51, doi:10.1126/science.1058040
- Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis, 2nd edition. *Meas-Interdiscip Res* 2019;17(3):160-167, doi:10.1080/15366367.2019.1565254
- Vorontsova D, Datsenko KA, Medvedeva S, Bondy-Denomy J, Savitskaya EE, Pougach K, Logacheva M, Wiedenheft B, Davidson AR, Severinov K, et al. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. *Nucleic Acids Res.* 2015 Dec 15;43(22):10848–60.
- Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. *Nucleic Acids Res.* 43, 10848–10860 (2015).
- Wakefield, N., R. Rajan, and E.J. Sontheimer, Primary processing of CRISPR RNA by the endonuclease Cas6 in *Staphylococcus epidermidis*. *FEBS Lett*, 2015. 589(20 Pt B): p. 3197-204.

Walker, M. W. G., Klompe, S. E., Zhang, D. J. & Sternberg, S. H. Transposon mutagenesis libraries reveal novel molecular requirements during CRISPR RNA-guided DNA integration. <http://biorxiv.org/lookup/doi/10.1101/2023.01.19.524723> (2023) doi:10.1101/2023.01.19.524723.

Wang JC. Helical repeat of DNA in solution. *Proc Natl Acad Sci.* 1979 Jan 1;76(1):200–3.

Wang, J. et al. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* 163, 840–853 (2015).

Wang, J. Y., Pausch, P. & Doudna, J. A. Structural biology of CRISPR–Cas immunity and genome editing enzymes. *Nature Reviews Microbiology* 0123456789, (2022).

Wang, R., Li, M., Gong, L., Hu, S. & Xiang, H. DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in *Haloarcula hispanica*. *Nucleic Acids Res.* 44, 4266–4277 (2016).

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., and Tan, W. (2020). Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*.

Wang, X. et al. Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. *Nat. Struct. Mol. Biol.* 23, 868–870 (2016).

Wang, Y., et al., Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*, 2009. 461(7265): p. 754-61.

Wang, Y., et al., Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 2008. 456(7224): p. 921-6.

Watts, E. A. et al. Histones direct site-specific CRISPR spacer acquisition in model archaeon. *Nat Microbiol* (2023) doi:10.1038/s41564-023-01446-3.

Wei Y, Chesne MT, Terns RM, Terns MP. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res.* 2015 Feb 18;43(3):1749–58.

Wenck, B. R. & Santangelo, T. J. Archaeal transcription. *Transcription* 11, 199–210 (2020).

Westra ER, van Houte S, Oyesiku-Blakemore S, Makin B, Broniewski JM, Best A, Bondy-Denomy J, Davidson A, Boots M, Buckling A. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr Biol.* 2015 Apr;25(8):1043–9.

Westra, E. R. et al. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr. Biol.* 25, 1043–1049 (2015).

Whitfield, C. R., Wardle, S. J. & Haniford, D. B. The global bacterial regulator H-NS promotes transpososome formation and transposition in the Tn5 system. *Nucleic Acids Res.* 37, 309–321 (2009).

WHO (2019). SITUATION REPORT - 1.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemond G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *J Open Source Softw.* 2019 Nov 21;4(43):1686.

Wickham H. *ggplot2*. Cham: Springer International Publishing; 2016.

Wiedenheft, B. et al. Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. *Structure* 17, 904–912 (2009).

Wiedenheft, B., et al., RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci U S A*, 2011. 108(25): p. 10092-7.

Wiegand T, Semenova E, Shiriaeva A, Fedorov I, Datsenko K, Severinov K, Wiedenheft B. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. *Cris J.* 2020 Oct 1;3(5):378–87.

Wiegand, T. et al. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. *CRISPR J.* 3, 378–387 (2020).

Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation: PROTEIN SCIENCE.ORG. *Protein Sci.* 27, 293–315 (2018).

Willkomm S, Makarova KS, Grohmann D. DNA silencing by prokaryotic Argonaute proteins adds a new layer of defense against invading nucleic acids. *FEMS microbiology reviews.* 2018;42(3):376-87.

Willkomm S, Oellig CA, Zander A, Restle T, Keegan R, Grohmann D, et al. Structural and mechanistic insights into an archaeal DNA-guided Argonaute protein. *Nature Microbiology.* 2017;2:17035.

Wright A V., Doudna JA. Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol.* 2016 Oct 5 [cited 2017 Dec 9];23(10):876–83.

Wright A V., Liu J-J, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the CRISPR genome integration complex. *Science* (80-). 2017 Sep 15;357(6356):1113–8.

Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* 23, 876–883 (2016).

- Wright, A. V. et al. Structures of the CRISPR genome integration complex. *Science* 357, 1113–1118 (2017).
- Wu Z, Tan S, Xu L, Gao L, Zhu H, Ma C, et al. NgAgo-gDNA system efficiently suppresses hepatitis B virus replication through accelerating decay of pregenomic RNA. *Antiviral research*. 2017;145:20-3.
- Wu, F., Xiao, A., Zhang, J., Gu, X., Lee, W.L., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N., Endo, N., et al. (2020). SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. *MedRxiv*.
- Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265-269.
- Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., Yin, H., Xiao, Q., Tang, Y., Qu, X., et al. (2020). Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *Lancet Gastroenterol Hepatol*.
- Wurtzer, S., Marechal, V., Mouchel, J.-M., Maday, Y., Teyssou, R., Richard, E., Almayrac, J.L., and Moulin, L. (2020). Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. *medRxiv*.
- Xiao Y, Ng S, Nam KH, Ke A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature*. 2017 Oct 4;550(7674):137–41.
- Xiao, F., Sun, J., Xu, Y., Li, F., Huang, X., Li, H., Zhao, J., Huang, J., and Zhao, J. (2020). Infectious SARS-CoV-2 in Feces of Patient with Severe COVID-19. *Emerg Infect Dis* 26.
- Xiao, Y. et al. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. 0839, 1–12 (2018).
- Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* 550, 137–141 (2017).
- Xu, Y.e.a. (2020). Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nat Med*.
- Yair, Y. & Gophna, U. Repeat modularity as a beneficial property of multiple CRISPR-Cas systems. *RNA Biology* 16, 585–587 (2019).
- Ye, Q. et al. HORMA Domain Proteins and a Trip13-like ATPase Regulate Bacterial cGAS-like Enzymes to Mediate Bacteriophage Immunity. *Mol Cell* 77, 709-722.e7 (2020).

Ye, Y., Ellenberg, R.M., Graham, K.E., and Wigginton, K.R. (2016). Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater. *Environmental Science & Technology* 50, 5077-5085.

Yoganand KNR, Sivathanu R, Nimkar S, Anand B. Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res.* 2017 Jan 9;45(1):367–81.

Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* 2012;40(12):5569–76.

Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* 40, 5569–5576 (2012).

Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Battistuzzi FU, editor. *Mol Biol Evol.* 2018 Dec 1;35(12):3041–3.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution.* 2017;8(1):28-36.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. Package for visualization and annotation of phylogenetic trees with their covariates and other associated data. McInerny G, editor. *Methods Ecol Evol.* 2017 Jan 22;8(1):28–36.

Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinforma.* 2020 Mar 5;69(1).

Yu, G., Lam, T. T. Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* 35, 3041–3043 (2018).

Yu, G., Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics*, 2020. 69(1): p. e96.

Yuan Y-R, Pei Y, Ma J-B, Kuryavyi V, Zhadina M, Meister G, et al. Crystal Structure of *A. aeolicus* Argonaute, a Site-Specific DNA-Guided Endoribonuclease, Provides Insights into RISC-Mediated mRNA Cleavage. *Molecular Cell.* 2005;19(3):405-19.

Zamore, P.D., RNA interference: big applause for silencing in Stockholm. *Cell*, 2006. 127(6): p. 1083-6.

Zander A, Willkomm S, Ofer S, van Wolferen M, Egert L, Buchmeier S, et al. Guide-independent DNA cleavage by archaeal Argonaute from *Methanocaldococcus jannaschii*. *Nature Microbiology.* 2017;2:17034.

Zayed, H. The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. *Nucleic Acids Res.* 31, 2313–2322 (2003).

Zehender, G., Lai, A., Bergna, A., Meroni, L., Riva, A., Balotta, C., Tarkowski, M., Gabrieli, A., Bernacchia, D., Rusconi, S., et al. (2020). Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J Med Virol.*

Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics.* 2014 Mar 1;30(5):614–20

Zhu, W. et al. The CRISPR ancillary effector Can2 is a dual-specificity nuclease potentiating type III