Curiositas

Journal of undergraduate research at Montana State University Fall 2021

Title:
Bioinformatic Identification of CRISPR Leader Motifs

Authors:
Pushya Krishna

Author Affiliation:
Department of Microbiology & Cell Biology, Montana State University

Volume
Fall 2021

Pages:
14-18

Abstract:
Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins serve as a sophisticated adaptive immune system to defend bacteria and archaea from viral infection. CRISPR mediated immunity occurs in three stages which allow the bacteria to adapt and respond to new as well as previously encountered viruses. The initial step of CRISPR adaptation requires the help of the Integration Host Factor (IHF) and a stretch of 200 base pairs known as the CRISPR leader to ensure encounters with new viruses are properly recorded in the host organism's immunological memory. A bioinformatic analysis of over 15,000 CRISPR leaders reveals that IHF is a prevalent and widespread feature of CRISPR adaptation across several different CRISPR subtypes and host organisms.

Curiositas is an interdisciplinary research journal dedicated to presenting the breadth and depth of undergraduate research that occurs at Montana State University. The journal places a particular emphasis on showcasing overlooked domains of undergraduate research, such as the humanities and arts, alongside traditional scientific research. *Curiositas* is committed to the belief that research on MSU's campus does not just occur in large laboratories and research groups: it occurs in every discipline and touches every element of scholarship that occurs at MSU. Articles in *Curiositas* are reviewed by a faculty member in the appropriate discipline (where applicable) and by an interdisciplinary undergraduate review committee.

Montana State University
Department of Microbiology & Cell Biology
109 Lewis Hall · PO Box 173520 · Bozeman, MT 59717
406-994-2902 · mcb@montana.edu

Please send questions and comments to
CuriositasJournal@montana.edu

montana.edu/curiositas

Microbiology & Cell Biology
http://doi.org/10.15788.f2021.curio3

Bioinformatic Identification of CRISPR Leader Motifs

Pushya Krishna
Department of Microbiology & Cell Biology, Montana State University

Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins serve as a sophisticated adaptive immune system to defend bacteria and archaea from viral infection. CRISPR mediated immunity occurs in three stages which allow the bacteria to adapt and respond to new as well as previously encountered viruses. The initial step of CRISPR adaptation requires the help of the Integration Host Factor (IHF) and a stretch of 200 base pairs known as the CRISPR leader to ensure encounters with new viruses are properly recorded in the host organism's immunological memory. A bioinformatic analysis of over 15,000 CRISPR leaders reveals that IHF is a prevalent and widespread feature of CRISPR adaptation across several different CRISPR subtypes and host organisms.

Introduction:

Perhaps now more than ever, humanity has become intimately familiar with the threat viruses pose to its survival. Since 2019, the SARS-Cov-2 virus has spread through communities across the globe and has established itself as one of the deadliest pandemics in human history. However, while the recent pandemic has reinvigorated public interest in the ubiquitous threat of viruses, bacteria and archaea have been quietly waging a constant biological war with their viral counterparts for billions of years. Known as bacteriophage, these viruses infect bacteria and archaea at a staggering rate of over 10^{23} infections per second.[1] As a result, this constant biological conflict has forced bacteria and archaea to evolve sophisticated immune systems to defend themselves against viral infection. One of the most well-known immune systems is the CRISPR adaptive immune system. CRISPR has gained significant spotlight in the public eye for its potential as a tool for genetic modification. However, its properties as a microbial immune system remain equally fascinating and is the focus for this article.

The Anatomy of a CRISPR

Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins serve as a sophisticated adaptive immune system to protect bacteria and archaea from viral infection. The acronym CRISPR describes a distinct genomic signature that serves as the focal point for these microorganism's immunological memory.[2] As seen in Figure 1a, a CRISPR consists of an array of identical repeating sequence elements (black squares) interspersed by sections of DNA that have been derived from previous viral infections (colored diamonds) (Figure 1a). The CRISPR array is accompanied by a 200 base pair upstream regulatory region known as the CRISPR leader[3] and a nearby region known as the Cas operon, which encodes the genes that produce Cas proteins.[2] Collectively, the CRISPR array, CRISPR leader and Cas operon are referred to as the CRISPR locus. Interestingly, CRISPR systems occur in immense diversity and are categorized into separate "subtypes" based upon the assortment of Cas proteins encoded by the Cas operon.[4]

CRISPR Adaptive Immunity Occurs in Three Stages

A key feature of CRISPR mediated immunity is the ability for the microorganism to "remember" which viruses have previously infected the cell. This allows the bacteria or archaea to prepare a unique self-defense for each virus that it encounters. The process by which CRISPR systems help defend microorganisms occurs in three distinct stages.[2] The first stage, known as CRISPR adaptation, occurs when a bacteria or archaea is infected by a new virus. During this novel encounter, the host organism will excise and insert a portion of the viral genome into the CRISPR locus (Figure 1b). Since the viral DNA is inserted in

between a pair of CRISPR repeats, the inserted sequences are referred to as "spacers". A key feature of CRISPR adaptation is that it allows the host organism to maintain a biological record of all previous viral infections. However, to convert this genetic record book into an immune response for self-defense, a second stage known as "expression and processing" must take place. In the expression phase, each spacer within the CRISPR array is converted from DNA into an individual RNA molecule known as a CRISPR RNA (crRNA). At the same time, the genes within the Cas operon are transcribed and translated into Cas proteins that function as the immunological workhorses for the cell. Certain Cas proteins have an extraordinary ability to cut DNA in a highly specific manner. However, they require assistance from the crRNA to successfully identify a genetic target. During the crRNA processing phase, the crRNA and certain Cas proteins are combined into a complex that can recognize and cut specific genetic targets (Figure 1c). Once they are successfully formed, these crRNA-Cas protein complexes can then perform the third stage of CRISPR immunity known as interference. During the life cycle of the cell, the crRNA-Cas protein complex swims around the interior of the cell in search of a viral genome. The crRNA functions as a biological search warrant and when a viral genome that matches the crRNA is detected, the crRNA-Cas protein complex can begin cutting apart the viral DNA and successfully protect the host organism from an infection[2] (Figure 1d).

The ability for crRNA-Cas protein complexes to identify and cut sequences of DNA in a highly specific manner has resulted in these microbial immunological devices being repurposed for a variety of applications. Recently, CRISPR based technologies have been used to treat genetic disorders such as sickle-cell anemia[5] and cystic fibrosis[6]. Alternatively, the pandemic has motivated scientists to repurpose CRISPR based technologies as a diagnostic tool for viruses such as SARS-Cov2. The discovery of CRISPR adaptive immunity has both improved our understanding of how microorganisms defend themselves against viral infection as well as kickstarted an entire decade of unparalleled biomedical innovation to treat diseases previously considered incurable.

Why is the CRISPR Leader Important?

It has been shown that new viral DNA is preferentially introduced at the "leader" end of the CRISPR array rather than being incorporated at a random location. The answer for why this distinct pattern occurs during CRISPR adap-

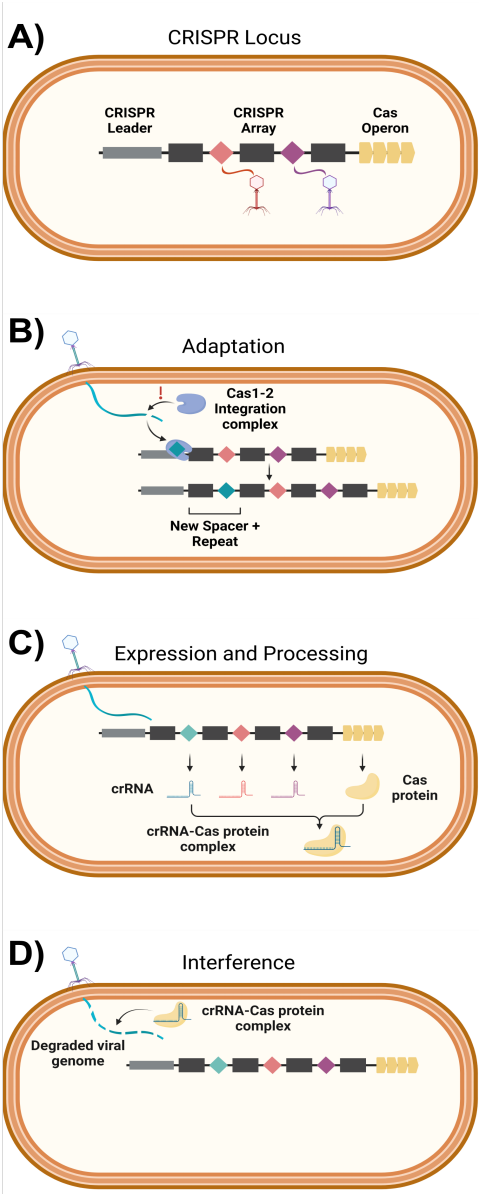

FIGURE 1

CRISPR overview

A) Diagram of a CRISPR locus containing a CRISPR leader, CRISPR array and Cas operon; B) "leader-end" CRISPR adaptation; C) CRISPR expression and processing; D) CRISPR interference

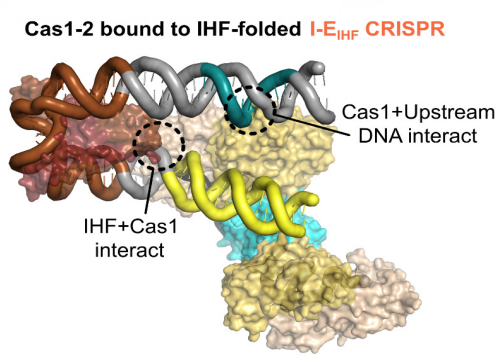

FIGURE 2

Cas1-2 bound to an IHF-folded CRISPR leader from Type I-E CRISPR in *E. coli*.

IHF binding to its respective binding site (brown) results in a 180° bend in the Leader DNA. DNA bending results in an upstream motif (turquoise) being recruited to the Cas1-2 complex located at the leader-repeat junction (yellow).

tation is contained within the CRISPR leader. The Doudna lab, lead by Nobel Prize-winning CRISPR pioneer Jennifer Doudna, demonstrated that in the Type I-E CRISPR locus of *Escherichia coli* (K12), a protein known as the Integration Host Factor (IHF) binds to a specific sequence of DNA in the CRISPR leader. Upon binding, IHF kinks the CRISPR leader, creating a horseshoe-shaped structure that stabilizes the protein complex which carries out adaptation, known as the Cas1-2 complex.[7] IHF-induced kinking of the CRISPR leader ensures that CRISPR adaptation occurs at the "leader" end of the CRISPR array by trapping the Cas1-2 integration complex in the curve of the horseshoe. Additionally, DNA kinking recruits a unique sequence of DNA known as an upstream-sequence motif (UM) that comes in direct contact and further stabilizes the Cas1-2 integration complex[8] (Figure 2).

Featured Academic Contribution: Identification of Novel CRISPR Leader Motifs (Santiago-Frangos et al., 2021)

The work performed by the Doudna lab has led to a paradigm for understanding the mechanism of polarized CRISPR evolution. However, we recently noticed that the IHF binding site in the leader sequence of the type I-F CRISPR system in *Pseudomonas aeruginosa* (PA14) is 8 base pairs further from the repeat sequence than originally observed in the type I-E systems from E. *coli* (K12). This seemingly subtle difference has important mechanistic implications.

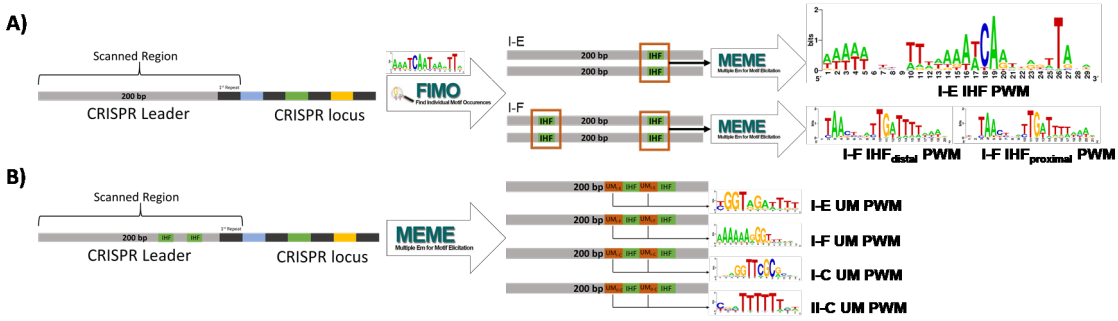

FIGURE 3

Schematic of the bioinformatic pipeline designed to identify conserved sequence motifs in CRISPR leaders.

A) Leader sequences were scanned to identify IHF binding sites using a tool known as FIMO. Matched regions were subsequently analyzed via MEME to develop a position weight matrix (PWM) of the IHF binding site. In the PWM, the size of each nucleotide is proportional to its relative frequency at each position. B) The search for a conserved UM was localized to leader regions containing a detected IHF motif. Leader sequences were scanned with MEME resulting in UM PWM's for I-E, I-F, I-C and II-C leaders.

The helical structure of double stranded DNA (dsDNA), made famous by Watson and Crick, contains 10.5 base pairs per 360-degree rotation. Thus, the addition of 8 base pairs not only shifts the IHF binding site a total 24 Å from the repeat but also introduces a ~260° rotation. Since IHF is a DNA kinking protein that recruits upstream DNA to a specific location on the Cas1-2 integration machinery, this 8 base pair insertion will reposition the upstream motif to a distinct location on integration machinery in *P. aeruginosa*. The unique leader architecture of *P. aeruginosa* led me to hypothesize that the distribution and composition of IHF binding sites and upstream motifs varies in a CRISPR subtype specific manner.

To test this hypothesis, I worked alongside a team of bioinformaticians in Dr. Blake Wiedenheft's lab to develop a bioinformatic pipeline to analyze 15,274 CRISPR leader sequences for the presence of conserved IHF binding sites and upstream motifs (Figure 3). The resulting data lead to four key discoveries. First, IHF binding sites are not restricted to I-E and I-F systems. Instead, IHF binding sites were found in 59% of I-C, 28% of I-E, 67% of I-F and 44% of II-C leaders, suggesting that IHF mediated integration is more widespread than previously understood. 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. Second, as initially observed in the PA14 strain of P. aeruginosa, the vast majority of I-F leaders possess an IHF motif situated further upstream of the first repeat relative to IHF binding sites in I-E leaders (Figure 4). Third, highly conserved upstream motif sequences were identified in 66% of I-C, 77% of I-E, 87% of I-F and 33% of II-C IHF-containing leaders. This suggests a widespread IHF-mediated integration mechanism similar to the previously established model in E. coli. Although different CRISPR subtypes (i.e. I-E, I-F, I-C etc.) utilize different Cas proteins during CRISPR immunity, it appears that they rely on similar mechanisms to regulate CRISPR adaptation. Fourth, we discovered that naturally occurring insertions and deletions displaced the position of a given motif found in different leaders by 9-12 base pairs (roughly a helical turn). This analysis provided bioinformatic evidence that there is a selective pressure to preserve the phased display of these motifs.

Collectively, these data demonstrate two key points that modify the previously established paradigm for leader-end evolution of CRISPR loci and support my initial hypothesis. First, IHF-dependent CRISPR integration mechanisms are more widespread than previously understood and second, leader motif architectures are varied across subtypes.

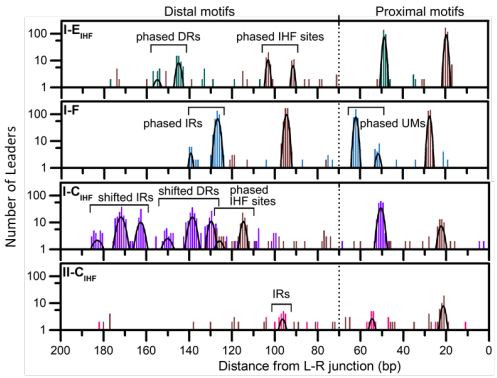

FIGURE 4

Subsets of I-E, I-F, I-C and II-C leaders display unique patterns of phased conservation of IHF and UM sequences.

Peaks of motif occurrence are plotted along the length of the leader. Each peak represents instances of motif occurrence at specific loactions along the length of the CRISPR leader.

Pushya Krishna is a senior double majoring in Cell Biology and Neuroscience and English Literature. Pushya started his research career in the Schmidt lab studying the impact of oxidative stress on liver cancer development in mouse models. Since 2018, Pushya has been working in the Wiedenheft lab where he has transitioned to studying the regulation and function of CRISPR adaptive immune systems. His areas of interest include bioinformatics, antibiotics and antimicrobial growth agents and CRISPR based diagnostic technologies. Originally from Bozeman, when he is not in the lab, Pushya can be frequently found playing soccer, making music with the violin or the flute or spending time with his family.

To learn more about the conclusions of this research, please find our lab's article published in *Current Biology*: **DOI 10.1016/j.cub.2021.05.068**

Acknowledgements

I would like to thank Dr. Blake Wiedenheft, Dr. Andrew Santiago-Frangos, Tanner Wiegand and Murat Buyukyoruk for their contributions to this research. Research reported in this presentation was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103474 to B.B. and R35 GM134867 NIGMS MIRA to B.W.

References

1.Microbiology by numbers. Nature Reviews Microbiology. 2011;9(9):628-.

2.Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007;315(5819):1709-12.

3.Alkhnbashi OS, Shah SA, Garrett RA, Saunders SJ, Costa F, Backofen R. Characterizing leader sequences of CRISPR loci. Bioinformatics. 2016;32(17):i576-i85.

4.Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015;1311:47-75.

5.Frangoul H, Altshuler D, Cappellini MD, Chen Y-S, Domm J, Eustace BK, et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. New England Journal of Medicine. 2020;384(3):252-60.

6.Maule G, Casini A, Montagna C, Ramalho AS, De Boeck K, Debyser Z, et al. Allele specific repair of splicing mutations in cystic fibrosis through AsCas12a genome editing. Nature Communications. 2019;10(1):3556.

7.Nuñez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature. 2015;527(7579):535-8.

8.Wright AV, Liu J-J, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the CRISPR genome integration complex. Science. 2017;357(6356):1113-8.