STRUCTURAL CHARACTERIZATION OF THE CSA3/CA_4 COMPLEX - A NEXUS FOR CLASS 1 CRISPR-CAS IMMUNE RESPONSE COORDINATION & ESTABLISHING A CURE FOR HIGHLY EFFICIENT GALECTIN EXPRESSION

by

Alexander Anthony Charbonneau

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

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2024

©COPYRIGHT

by

Alexander Anthony Charbonneau

2024

All Rights Reserved

ACKNOWLEDGEMENTS

The achievement represented by this dissertation would not have been possible without the help and support of numerous mentors, teachers, co-workers, collaborators, and family members. Particular thanks to my advisor, Dr. Martin Lawrence for his guidance, to Colin Gauvin for his aid in many experiments, Dr. Mary Cloninger for her help managing a team of undergraduates, and to my wife, Kasey Kane Charbonneau, for her unending support throughout my journey here.

TABLE OF CONTENTS

1. INTRODUCTION 1

Virus-Host Interactions 1
The CRISPR-Cas Immune System 1
CRISPR-Cas is a Highly Diverse Immune System 6
Class 1 CRISPR-Cas Systems 11
Ancillary CRISPR-Cas Proteins and CARF Domains 16
Unique CRISPR-Cas Second Messengers: Cyclic Oligo-Adenylates 17
The role of Csa3a in CRISPR-Cas 19
Csa3a and cA_4: a Nexus for Class 1 Systems 19
High Level Protein Expression For the Study of Human Galectins and Undergraduate Research Pursuits 20
References 22

2. METHODS AND METHODOLOGY 30

Expanded Methods for Manuscript 1 30
Bioinformatics 30
Sulfolobus Strains, Growth Conditions 30
Csa3a Expression and Purification in Sulfolobus Solfataricus 31
Screening for Csa3a-cA_4 Crystallization Conditions 32
Quantitation of EMSA Binding Assays 33
References 34

3. CYCLIC TETRA-ADENYLATE (CA_4) RECOGNITION BY CSA3; IMPLICATIONS FOR AN INTEGRATED CLASS 1 CRISPR-CAS IMMUNE RESPONSE IN *S. SOLFATARICUS* 36

Contribution of Authors and Co-Authors 36
Manuscript Information 37
Abstract 38
1.1 CARF Domains 40
1.2 Cyclic Oligo-Adenylate Second Messengers 41
1.3 Transcriptional Regulation by Csa3 41
1.4 Integration of Class 1 Systems 42
2. Materials and Methods 43
2.1 Bioinformatics 43
2.2 Csa3a Expression and Purification 43
2.3 Isothermal Titration Calorimetry 45
2.4 Cyclic Oligonucleotide Phosphodiesterase (Ring Nuclease) Assays 45

TABLE OF CONTENTS CONTINUED

2.5 Crystallization and Data Collection 46
2.6 Structure Determination and Refinement 47
2.7 Electromobility Shift Assays (EMSAs) 49
2.8 Fluorescence Polarization (FP) Assays 50
2.9 Sulfolobus Strains, Growth Conditions 51
2.10 Acquisition Assay 51
3. Results 51
3.1 CRISPR/Cas Systems in S. solfataricus 51
3.2 CRISPR Associated Palindromes 53
3.3 Csa3a Binds cA_4 54
3.4 Csa3a Lacks Ring Nuclease Activity 56
3.5 Csa3a Binds cA_4 in the CARF Binding Pocket 57
3.6 cA_4 Induced Conformational Changes 62
3.7 Csa3a Binds CAPPa Nonspecifically, in a Cooperative and cA_4 Independent Manner 65
3.8 Csa3a Overexpression Does Not Stimulate Spacer Acquisition in S. solfataricus 69
4. Discussion 71
4.1 Physiological Implications of Low Micromolar cA_4 Affinity 71
4.2 cA_4 Induced Conformational Changes 72
4.3 cA_4 Recognition is Entropy Driven 72
4.4 An Integrated, Synergistic Class 1 CRISPR-Cas Immune Response 76
4.5 Caveats and Additional Considerations 77
References 78

4. A CURE FOR HIGH LEVEL GALECTIN EXPRESSION 84

Contribution of Authors and Co-Authors 84
Manuscript Information 87
Abstract 88
1. Introduction 89
2. Materials and Methods 92
2.1. Galectin-3 Expression Clones 92
2.2. Galectin-7 and Galectin-8 Expression Clones 93
2.3. IPTG Induction 93
2.4. Autoinduction 94
2.5. ZYP-5052 Autoinduction Media 94
2.6. Purification 95
2.7. Analytical Size Exclusion Chromatography 96
2.8. Mass Spectroscopy 96
2.9. Fluorescence Lifetime Binding Assays 97

TABLE OF CONTENTS CONTINUED

3. Results 97
3.1. Low-Level Expression of Gal-3 CRD with IPTG 97
3.2. Basal Gal-3 CRD Expression in the Absence of IPTG 99
3.3. Gal-3 CRD Binds IPTG 100
3.4. Autoinduction Produces High Gal-3 CRD Yields 102
3.5. Autoinduction, Expression, and Purification of Additional Galectins 104
4.1. Comparison ofAautoinduction and IPTG-Induction for Galectin Expressions 105
4.2. A "CURE" for Galectin Expression, Purification, and Characterization 108
4.2.1. Overview and Comparison of CUREs and UREs 108
4.2.2. A CURE for Galectin Expression, Purification, and Characterization. 110
5. Conclusion 115
References 116

5. CSX3 IS A CYCLIC OLIGONUCLEOTIDE PHOSPHODIESTERASE ASSOCIATED WITH TYPE-III CRISPR-CAS THAT DEGRADES THE SECOND MESSENGER CA_4 121

Contribution of Authors and Co-Authors 121
Manuscript Information 122
Abstract 123
Introduction 124
Results 127
Csx3 Associates with Type-III CRISPR-Cas Systems 127
Conserved Sequence Motifs in Csx3 128
Csx3 Binds cA_4 with Nanomolar Affinity 129
The Csx3 Fold 131
Ring Nuclease Activity 136
Discussion 138
Defining Elements of the Cyclic oligoAdenylate Recognition Fold (CARF) 138
The C-terminal Tail 140
Recognition of cA_4 142
Structural Basis for Ring Nuclease Activity 144
Utility of Unlabeled cA_4 Ring Nuclease and Fluorescence Lifetime Binding Assays 146
Physiological Specificity for cA_4 146

TABLE OF CONTENTS CONTINUED

Experimental Procedures 148
Gene Annotations 148
Csx3 Expression and Purification 149
Fluorescence Lifetime Binding Assays 150
Crystallization and Structure Determination 151
Cyclic Oligonucleotide Phosphodiester-ase Assays 152
Data Availability 153
References 154

6. CONCLUSION 158

Conclusions from Chapter Three 158
Future Studies 159
Investigating Potential Protein-Protein Interactions with Csa3a. 159
Structural Characterization of Csa3b 160
Conclusions from Chapter Four 161
References 164

APPENDICES 166

APPENDIX A 167

Supporting Online Materials For: Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response 167
References 178

APPENDIX B 179

Supplementary Data: A CURE for High Level Galectin Expression 179
Gel-Based MS/MS Analysis – Orbitrap Fusion (Performed by the IDEA National Resource for Quantitative Proteomics at the University of Arkansas Medical School) 183
Cellular Concentration of Galectin-3 CRD from Leaky Expression 185
References 186

CUMULATIVE REFERENCES CITED 188

LIST OF TABLES

Table Page

1. Table 1. Data Collection Parameters 47
2. Table 2. Model Refinement Parameters 49
3. Table 1. Data Collection and Model Refinement 133
4. Supplemental Table S1 – ITC Values 173
5. Supplemental Table S2. PCR primers used to generate EMSA fragments. 173
6. Supplemental Table S3 - Fluorescence Polarization DNA Constructs 174
7. Supplemental Table S4 - Acquisition Assay PCR Primers 174
8. Supplemental Table S6. Correlation of Csa3a and CAPPa. Csa3 presence or absence correlates strongly with the presence or absence of CAPPa in the acquisition cassette (aCas). 178
9. Supplemental Table S7 - RNA Conformers for cA_4* 178
10. Table S1. PCR Primers for Galectin-3 and Galectin-3 CRD 180
11. Table S2. Measured extinction coefficients 181
12. Table S3. Galectin-3 CRD 0.1 mM IPTG Purification Summary 185
13. Table S4. Galectin-3 CRD Autoinduction Purification Summary 186
14. Table S5. Mean yields of Gal-3 CRD with various induction protocols. 186

LIST OF FIGURES

Figure Page

1. Figure 1. Overview of the CRISPR-Cas immune response. 5
2. Figure 2. Class 1 CRISPR-Cas systems. 9
3. Figure 3. Class 2 CRISPR-Cas systems. 11
4. Figure 4. Interference in Type I CRISPR-Cas systems. 13
5. Figure 5. Interference in Type III CRISPR-Cas systems. 15
6. Figure 1 ImageJ Analysis of EMSA Gels. 34
7. Figure 1. CRISPR Associated Palindromes in S. solfataricus. 53
8. Figure 2. ITC analysis of the Csa3-cA4 interaction. 56
9. Figure 3. Csa3a does not exhibit ring nuclease activity. 57
10. Figure 4. The Csa3a CARF domain binds cA4. 59
11. Figure 5. cA4 induced conformational changes. 63
12. Figure 6. Conserved residues in Csa3a wHTH domain. 64
13. Figure 7. Nonspecific binding to the Type I-A(2) cas4a promoter. 67
14. Figure 8. Csa3a binds CAPPa nonspecifically, in a cooperative and cA4 independent manner. 69
15. Figure 9. Overexpression of Csa3a in S. solfataricus does not lead to spacer acquisition. 70
16. Figure 10. Model for an integrated Class 1 immune response. 75
17. Fig. 1. A schematic representation of the A) prototype, B) tandem-repeat, and C) chimera galectins. 89
18. Fig. 2. Gal-3 CRD induction with 0.1 mM IPTG. 98

LIST OF FIGURES CONTINUED

Figure Page

19. Fig. 3. Leaky expression of Gal-3 CRD. 100

20. Fig. 4. Affinity of Gallectin-3 for IPTG..... 102

21. Fig. 5. Autoinduction increases Gal-3 CRD expression. 103

22. Fig. 6. Autoinduction outperforms IPTG. 104

23. Fig. 7. Autoinduction efficiently produces each class of human galectin. 105

24. Fig. 8. Feedback inhibition. 107

25. Figure 1. AfCsx3 binds cA4 with high affinity. 130

26. Figure 2. The AfCsx3 Fold. 135

27. Figure 3. The dynamic C-terminal tail. 136

28. Figure 4. Af Csx3 ring nuclease activity. 138

29. Figure 5. Structural definition of the CARF domain core. 141

30. Figure 6. RNA density in the 3WZI Csx3 structure accommodates cA4. 144

31. Supplementary Figure S1. Overview of CRISPR/Cas elements in S. solfataricus P2. 169

32. Supplemental Figure S2. cA4 omit map. 169

33. Supplemental Figure S3. Nonspecific binding to the I-A (1) cas4a promoter and CRISPER locus F leader. 170

34. Supplemental Figure S4. Nonspecific binding to CRISPR loci B and E leaders. 171

35. Supplemental Figure S5. Nonspecific binding to the CRISPR locus F leader and I-A(1) cas1 promoter. 172

LIST OF FIGURES CONTINUED

Figure Page

36. Supplemental Table S5. Acquisition cassette synteny for Sulfolobales containing Csa3a. 176

37. Fig. S1. Nucleotide sequences for Galectin-7 and Galectin-8 181

38. Fig. S2. Galectin-3 size exclusion chromatography..... 181

39. Fig. S3. Galectin-3 CRD size exclusion chromatography..... 182

40. Fig. S4. Galectin 7 size exclusion chromatography..... 182

41. Fig. S5. Galectin-8 elutes as a monomer via size exclusion chromatography..... 183

42. Fig. S6. Gal-3 CRD MS/MS results..... 184

43. Fig. S7. Galectin-7 MS/MS results..... 185

44. Fig. S8. Galectin-8 MS/MS results..... 185

ABSTRACT

Though Class I CRISPR-Cas systems, primarily Type I and Type III, are the most abundant CRISPR systems in archaea and bacteria, mechanisms driving their immune response regulation are not well understood. Csa3 family transcription factors, composed of N-terminal CARF and C-terminal winged helix-turn-helix domains, are frequently encoded within Type I CRISPR-Cas systems. Csa3 transcription factors are hypothesized to bind cyclic oligoadenylate (cOA) second messengers produced by Type III interference complexes, likely modulating their DNA-binding activity. Therefore, we investigated the interaction between Csa3a and cyclic tetra-adenylate (cA_4). Isothermal titration microcalorimetry showed *S. solfataricus* Csa3a binds cA_4 at biologically relevant concentrations in an entropically driven interaction. Ring nuclease assays revealed Csa3a lacks self-regulatory phosphodiesterase activity exhibited by other CARF domain proteins. We crystallized and solved the structure of the Csa3/cA_4 complex, which revealed conserved motifs are responsible for cA_4 binding and illuminated significant conformational changes induced by the interaction. We also identified an 18-bp palindromic motif, which we designated CAPPa, that is conserved in the 27 sequenced members of the order Sulfolobales, and shows synteny with Csa3a and acquisition genes in these genomes. We found Csa3a binds CAPPa in a nonspecific, cooperative, and cA_4-independent manner. These characteristics suggest a more complex method of transcriptional regulation than previously hypothesized. However, the interaction between Csa3a and cA_4 confirmed here signifies a nexus between Type I and Type III systems; we thus propose a model in which this interaction coordinates the two arms of an integrated immune system to mount a synergistic, highly orchestrated, adaptive immune response.

We applied the workflow designed to produce significant protein quantities for crystallographic studies of Csa3a to the study of *Homo sapiens* galectin proteins, a family of ß-galactoside-binding proteins. Here, we identified a putative autoinhibitory mechanism affecting traditional IPTG-induction methods by characterizing IPTG-binding capabilities of galectins and quantifying basal protein expression over various IPTG concentrations. To bypass this predicted feedback loop, we employed a highly efficient and approachable autoinduction method, resulting in a 7-fold increase in protein expression. Much of this work was done in the context of a course-based undergraduate research experience with great success.

CHAPTER ONE

INTRODUCTION

Virus-Host Interactions

The relationship between viruses and host cells is often likened to an arms race, a constant cycle of evolution in which features of either side are altered to bypass defenses, evade recognition entirely, or engage in a counteroffensive. From a probability standpoint, viruses and other mobile genetic elements clearly have the upper hand as the numerically dominant biological entities on earth [1]. Cellular life rarely benefits from a spatial or statistical advantage; thus, prokaryotes have evolved systems to target nearly every pathway to infection and every stage of the infection cycle. Innate immune systems such as restriction-modification and DNA phosphorothioation [2, 3] offer protection to cells in addition to more passive defenses from adsorption-blocking systems like receptor blocking, obscuration by extracellular matrix production, and competitive inhibitor production [4-7]. Though these innate systems instill a modicum of protection, they lack the ability to adapt quickly to the rapid evolutionary rates that viruses achieve.

The CRISPR-Cas Immune System

Another layer of defense in many prokaryotic immune defenses comes in the form of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) that perform adaptive and heritable immune functions with CRISPR-associated (*cas*)

genes [8-15]. Present in many bacteria and most archaea, these repeats are separated by short "spacers" derived from viruses, phage or other mobile genetic elements that have previously infected the cell or cells in its direct parental lineage. Because the spacers are incorporated into the genome, they are heritable, providing immunity to daughter cells.

Since the role of CRISPR-Cas in immunity was first recognized at the turn of the century [16-25], the basic tenets of CRISPR-Cas adaptive immunity have been well established [26-29]. This defense system is unique in prokaryotes for its heritable and adaptive qualities that differentiate it from the more passive systems described above. The immune response progresses in three distinct stages, i) adaptation, ii) crRNA maturation and iii) interference (Figure 1).

First, in adaptation, the adaptation/integration complex made up of CRISPR-associated proteins Cas1, Cas2 and Cas4 binds to, excises, and integrates short segments of invading DNA (protospacers) into CRISPR loci. These CRISPR loci, a series of spacers incorporated from invading nucleic acids separated by short repeat sequences are the origin of the namesake Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). Importantly, the unique adaptability of CRISPR systems and the ability to pass immunity to daughter cells is a direct result of this adaptation stage of the CRISPR immune response. While the role of the adaption mechanism is decidedly to build a library of targetable infective agents for use by downstream stages of the immune response, it comes at the cost of a non-negligible chance for autoimmunity by acquiring spacers consisting of host DNA [30, 31].

Cells must employ several mechanisms for the regulation of adaptation to mitigate the risks for self-targeting. A notable example of this regulation is found in the form of primed adaptation, in which adaptation machinery remains inactive or even entirely absent until stimuli from viral infection, such as envelope stress, is sensed [32-35]. Exact regulatory mechanisms, particularly for transcriptional regulation of adaptation genes, in primed adaptation events have yet to be fully characterized, and are thus a prime candidate for investigation.

In the second stage of the immune response, the CRISPR-locus is transcribed into a long RNA transcript containing multiple spacers and repeats called pre-CRISPR RNA. This pre-CRISPR RNA is processed into single spacers-repeat segments called CRISPR-RNA (crRNA) by Cas proteins (Cas6 or Cas12) or cellular nucleases. These mature crRNAs serve as the core for the construction of interference complexes or are loaded into those lacking crRNA already to serve as guides for targeting complimentary nucleic acids [36-38].

The third stage of the CRISPR-cas immune response harbors the active target recognition and degradation activities. After crRNA matures during the second stage and Cas proteins assemble into nucleoprotein complexes with individual crRNA guides, these inference complexes survey the cell for invading nucleic acids. Extreme diversity exists in the exact mechanisms of target recognition and degradation, but all systems utilize crRNA as a template to identify a target. Since the spacers encoded in the crRNA were originally excised from invading nucleic acids, their sequences are capable of making complimentary base-

pairing interactions with the same virus or mobile genetic element upon reinfection. When this occurs, the interference complex initiates degradation of the target to clear the infection. During a primed adaptation event triggered by this immune response, additional spacers may be acquired as the immune response begins again at the adaptation stage.

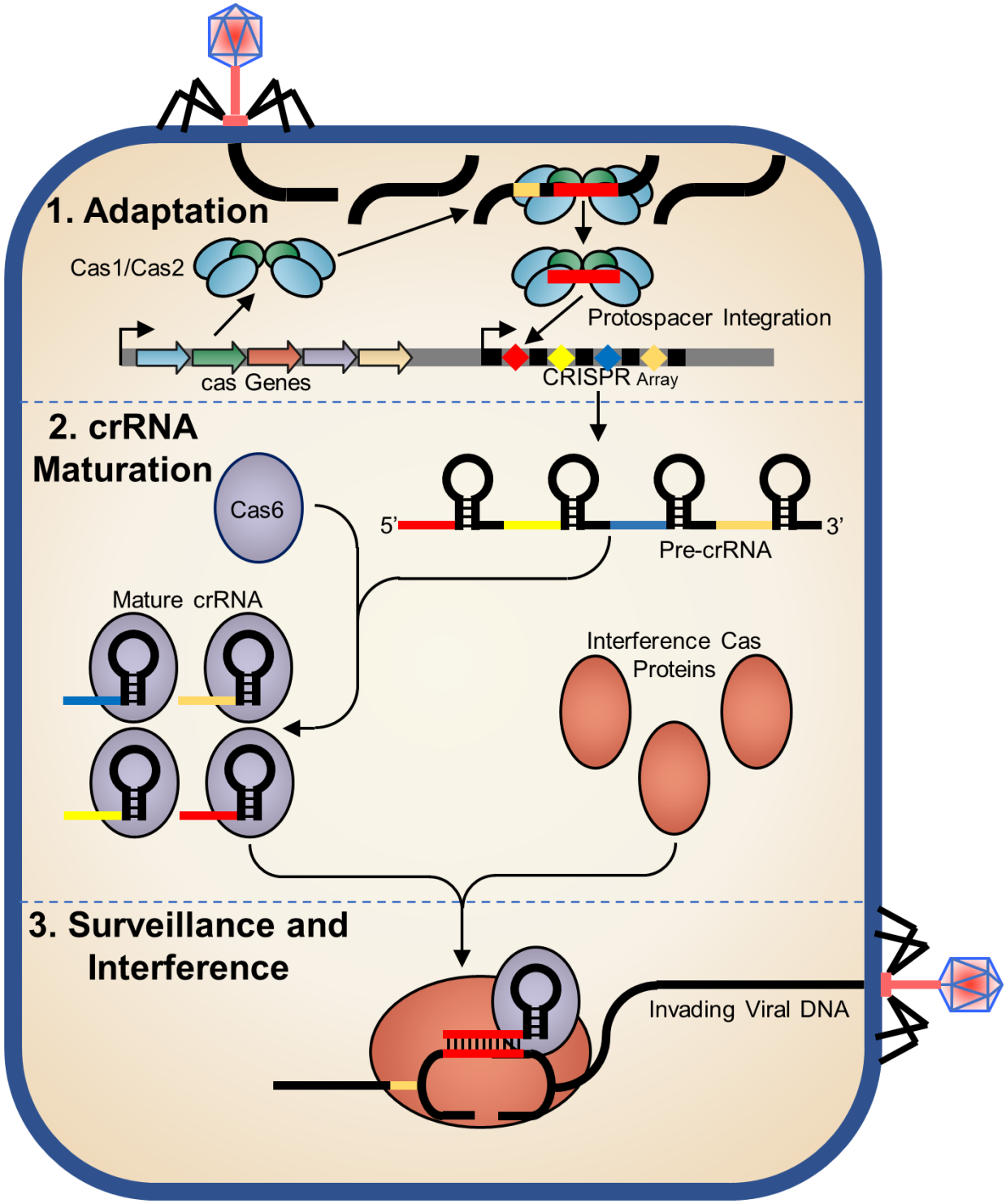


Figure 1. Overview of the CRISPR-Cas immune response. The CRISPR-Cas adaptive immune response is generally described in three stages, adaptation, crRNA maturation and interference. During adaptation, short DNA fragments from

invading nucleic acids are incorporated into CRISPR loci (diamonds) where they are separated by short palindromic DNA repeats (black rectangles). New spacers (red) flanked by PAM sequences (gold) are excised and inserted at the leader end of the CRISPR array by the Cas1-Cas2 integrase complex (blue/green), with assistance from additional type and subtype specific factors. In the second stage, crRNA maturation, the CRISPR array is transcribed as a single transcript of pre-crRNA. Type and subtype specific factors, often Cas6 (purple), cleave the transcript into individual mature CRISPR RNAs (crRNA). In stage 3, interference, crRNA and CRISPR associated (Cas) proteins assemble into interference complexes that survey the cell for invading nucleic acids, either RNA or DNA depending on the specific CRISPR-Cas system. When a complementary target or protospacer is identified, nucleases are activated to degrade the invading nucleic acid, eliminating the infection. Figure originally generated by the author and used with permission from Elsevier [39].

CRISPR-Cas is a Highly Diverse Immune System

CRISPR-Cas systems exhibit extreme diversity in form and function, and are first categorized into 2 distinct classes, Class 1 and Class 2, then further subdivided into 6 types and 33 different subtypes [13]. While diversity can be observed in all characteristics of the immune system, it is most apparent in the surveillance and interference machinery. In the broadest level of categorization, CRISPR-Cas systems are broken down into Class 1, systems which employ an interference complex consisting of multiple Cas proteins, and Class 2, systems that employ a single Cas protein in their interference complexes [13, 36, 37, 40-44]. These Classes are further divided into Types and Subtypes.

Class 1 systems are divided into Type I, Type III, and Type IV with Types I and III composing the majority of known Class 1 systems, leaving Type IV systems a rarity (Figure 2). Along with their abundance, Type I and Type III systems are often found encoded together within a single genome, sometimes in multiple copies

or subtypes [13]. While subtypes vary in their composition, Type I interference complexes generally contain Cas5, Cas6, multiple copies of Cas7, Cas8 and Cas11 in addition to crRNA. Since the Type I interference complex lacks the innate ability to cleave targeted DNA, it recruits Cas3 HD nuclease/helicase to degrade the invading nucleic acids. Type III surveillance complexes consist of many of the same Cas proteins including Cas5, multiple Cas7 subunits and Cas11, but also contain a Cas10 subunit with the ability to degrade targeted DNA.

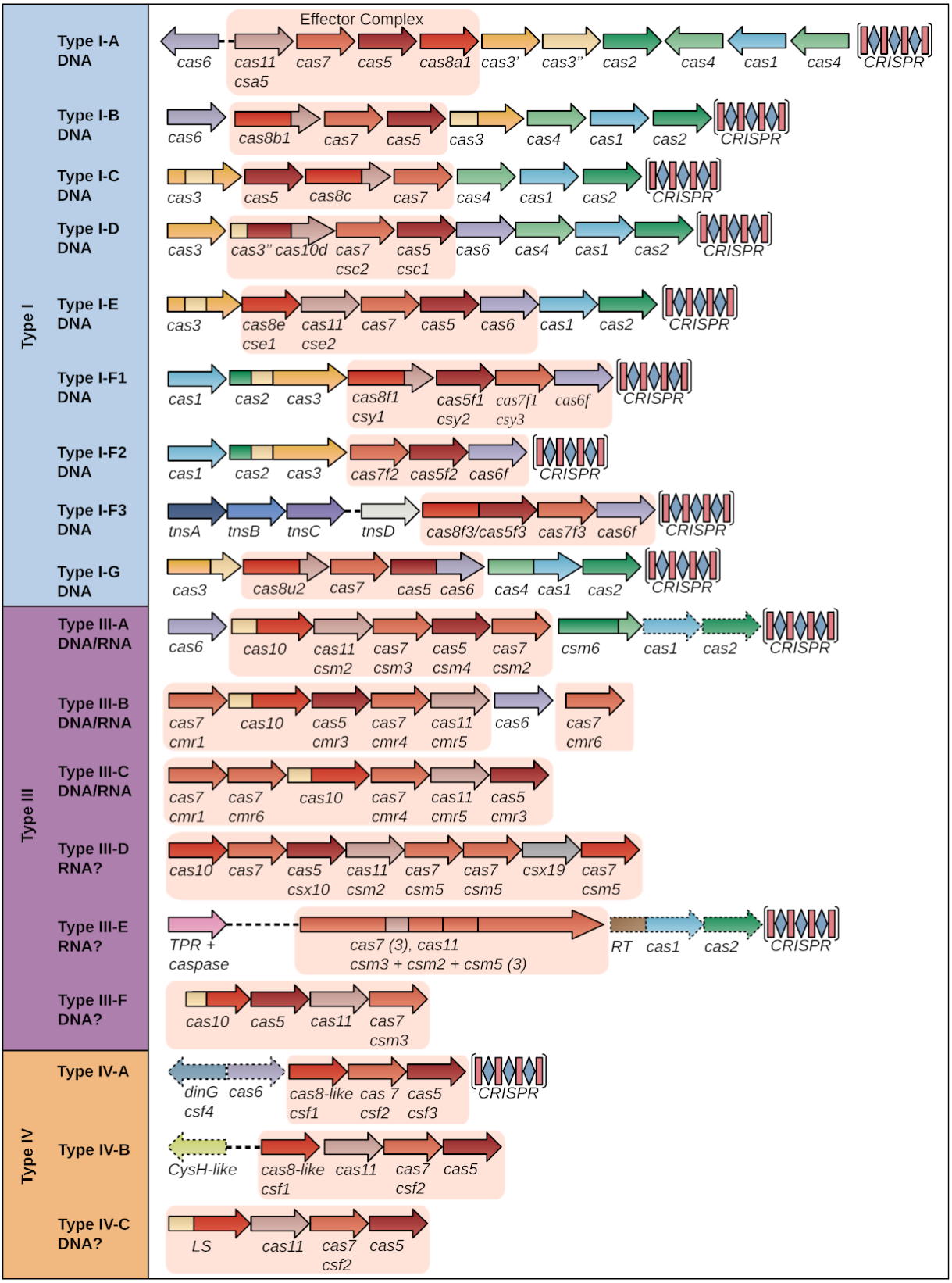

Type I
Type I-A DNA
Effector Complex
cas6
cas11 csa5
cas7
cas5
cas8a1
cas3'
cas3''
cas2
cas4
cas1
cas4
CRISPR
Type I-B DNA
cas6
cas8b1
cas7
cas5
cas3
cas4
cas1
cas2
CRISPR
Type I-C DNA
cas3
cas5
cas8c
cas7
cas4
cas1
cas2
CRISPR
Type I-D DNA
cas3
cas3'' cas10d
cas7 csc2
cas5 csc1
cas6
cas4
cas1
cas2
CRISPR
Type I-E DNA
cas3
cas8e cse1
cas11 cse2
cas7
cas5
cas6
cas1
cas2
CRISPR
Type I-F1 DNA
cas1
cas2
cas3
cas8f1 csy1
cas5f1 csy2
cas7f1 csy3
cas6f
CRISPR
Type I-F2 DNA
cas1
cas2
cas3
cas7f2
cas5f2
cas6f
CRISPR
Type I-F3 DNA
tnsA
tnsB
tnsC
tnsD
cas8f3/cas5f3
cas7f3
cas6f
CRISPR
Type I-G DNA
cas3
cas8u2
cas7
cas5 cas6
cas4
cas1
cas2
CRISPR
Type III
Type III-A DNA/RNA
cas6
cas10
cas11 csm2
cas7 csm3
cas5 csm4
cas7 csm2
csm6
cas1
cas2
CRISPR
Type III-B DNA/RNA
cas7 cmr1
cas10
cas5 cmr3
cas7 cmr4
cas11 cmr5
cas6
cas7 cmr6
Type III-C DNA/RNA
cas7 cmr1
cas7 cmr6
cas10
cas7 cmr4
cas11 cmr5
cas5 cmr3
Type III-D RNA?
cas10
cas7
cas5 csx10
cas11 csm2
cas7 csm5
cas7 csm5
csx19
cas7 csm5
Type III-E RNA?
TPR + caspase
cas7 (3), cas11 csm3 + csm2 + csm5 (3)
RT
cas1
cas2
CRISPR
Type III-F DNA?
cas10
cas5
cas11
cas7 csm3
Type IV
Type IV-A
dinG csf4
cas6
cas8-like csf1
cas 7 csf2
cas5 csf3
CRISPR
Type IV-B
CysH-like
cas8-like csf1
cas11
cas7 csf2
cas5
Type IV-C DNA?
LS
cas11
cas7 csf2
cas5

Figure 2. Class 1 CRISPR-Cas systems. Generalized CRISPR–cas loci of each Class 1 subtype and select subtype. Homologous genes are color-coded and identified by family name. CRISPR Loci are represented as a series of spacers and repeats (red rectangles and blue diamonds). Figure originally generated by the author and used with permission from Elsevier [39].

While this work focuses on Class 1 systems, it is important to know that Class 2 systems are likely more well-known due to their current application in genetic engineering and biotechnology applications. Class 2 systems are divided into Type II, Type V, and Type VI (Figure 3) with Type II systems utilizing the well-known Cas9 in their surveillance complex along with crRNA and tracrRNA. Like the Type II systems with their Cas9 cas protein, Type V systems utilize a single Cas12 unit and Type VI systems a single Cas13 unit [13].

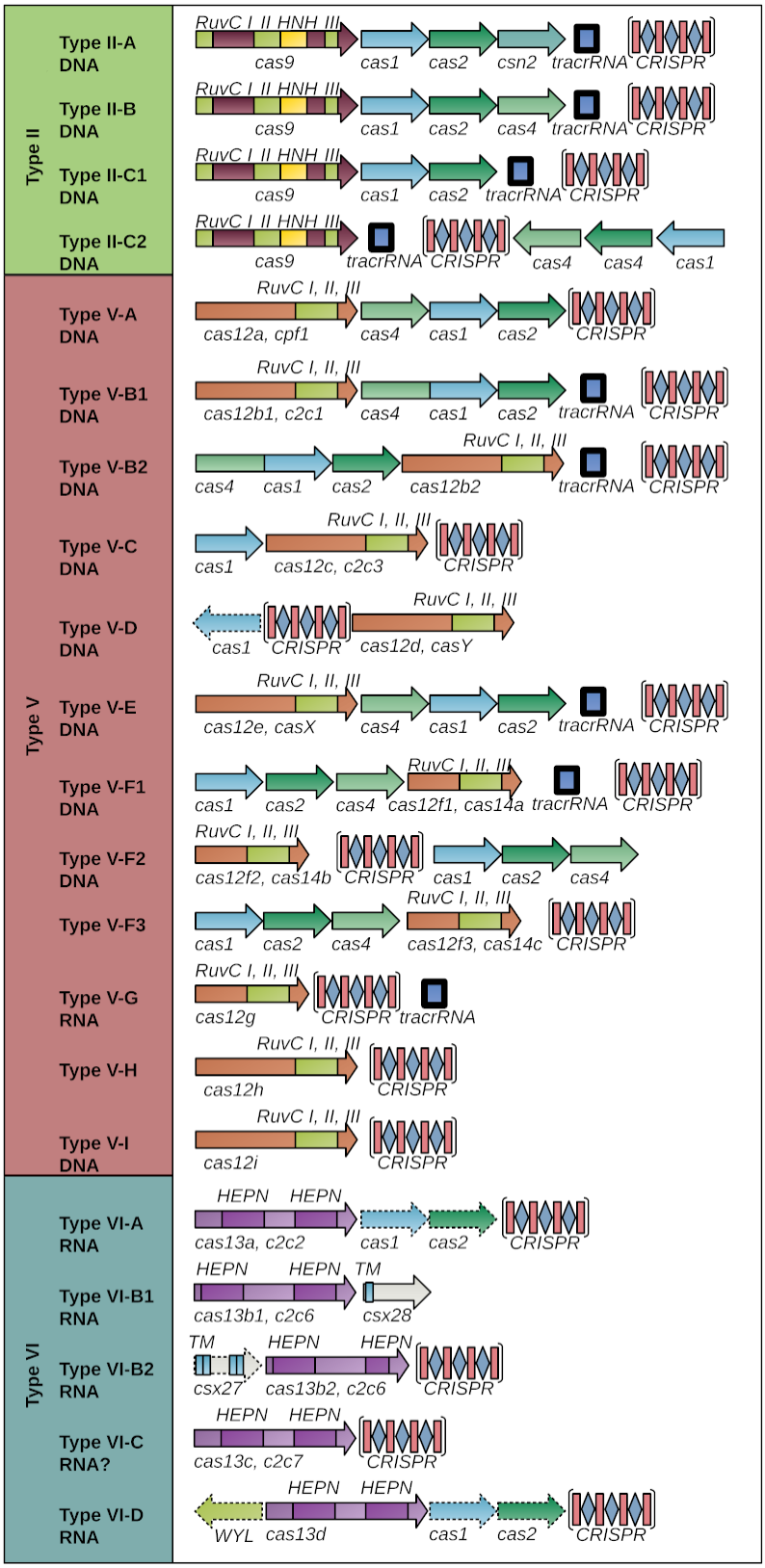

Type II
Type II-A DNA
RuvC I II HNH III
cas9
cas1
cas2
csn2
tracrRNA
CRISPR
Type II-B DNA
RuvC I II HNH III
cas9
cas1
cas2
cas4
tracrRNA
CRISPR
Type II-C1 DNA
RuvC I II HNH III
cas9
cas1
cas2
tracrRNA
CRISPR
Type II-C2 DNA
RuvC I II HNH III
cas9
tracrRNA
CRISPR
cas4
cas4
cas1
Type V
Type V-A DNA
RuvC I, II, III
cas12a, cpf1
cas4
cas1
cas2
CRISPR
Type V-B1 DNA
RuvC I, II, III
cas12b1, c2c1
cas4
cas1
cas2
tracrRNA
CRISPR
Type V-B2 DNA
cas4
cas1
cas2
RuvC I, II, III
cas12b2
tracrRNA
CRISPR
Type V-C DNA
cas1
RuvC I, II, III
cas12c, c2c3
CRISPR
Type V-D DNA
cas1
CRISPR
RuvC I, II, III
cas12d, casY
Type V-E DNA
RuvC I, II, III
cas12e, casX
cas4
cas1
cas2
tracrRNA
CRISPR
Type V-F1 DNA
cas1
cas2
cas4
RuvC I, II, III
cas12f1, cas14a
tracrRNA
CRISPR
Type V-F2 DNA
RuvC I, II, III
cas12f2, cas14b
CRISPR
cas1
cas2
cas4
Type V-F3
cas1
cas2
cas4
RuvC I, II, III
cas12f3, cas14c
CRISPR
Type V-G RNA
RuvC I, II, III
cas12g
CRISPR
tracrRNA
Type V-H
RuvC I, II, III
cas12h
CRISPR
Type V-I DNA
RuvC I, II, III
cas12i
CRISPR
Type VI
Type VI-A RNA
HEPN
HEPN
cas13a, c2c2
cas1
cas2
CRISPR
Type VI-B1 RNA
HEPN
HEPN
TM
cas13b1, c2c6
csx28
Type VI-B2 RNA
TM
csx27
HEPN
HEPN
cas13b2, c2c6
CRISPR
Type VI-C RNA?
HEPN
HEPN
cas13c, c2c7
CRISPR
Type VI-D RNA
HEPN
HEPN
WYL
cas13d
cas1
cas2
CRISPR

Figure 3. Class 2 CRISPR-Cas systems. Generalized CRISPR–cas loci of each Class 2 subtype and select subtype. Homologous genes are color-coded and identified by family name. CRISPR Loci are represented as a series of spacers and repeats (red rectangles and blue diamonds). Figure originally generated by the author and used with permission from Elsevier [39].

Class 1 CRISPR-Cas Systems

Relevant to the work described here are Class 1 CRISPR-Cas systems of Types/subtypes I-A, III-B, and III-D as these CRISPR-Cas systems are encoded in the *Saccharolobus solfataricus* P2 genome (formerly *Sulfolobus solfataricus* [45]). This model organism historically studied in our lab [46-51] encodes three Type I-A, three Type III-B, and one Type III-D systems, and harbors six CRISPR loci designated CRISPRs A–F (Supplemental Figure S1 in Appendix A) [52-54]. An important distinction between the immune response mechanisms of these CRISPR-Cas systems arises in their targeted nucleic acids and their method for identifying these targets.

Type I-A effector complexes (Figure 4), also known as archaeal Cascade (aCascade), survey cellular dsDNA for a protospacer-adjacent motif (PAM), a short 2-5 bp sequence proximal to the acquired spacer that [26, 37, 50]. The presence of a PAM proximal to the targeted protospacer and absence in the integrated spacer prevent these Type I systems from targeting the CRISPR array in an autoimmune mechanism. PAM recognition by the unique Cas8 subunit initiate local melting of the dsDNA, allowing the surveillance complex to interrogate the target for complimentary base pairing to the crRNA housed in the complex [55-57]. Notably,

the reliance on recognition of a PAM sequence to initiate target interrogation of degradation makes the Type I system vulnerable to mutations in the PAM or first few bases of the protospacer rendering an invading nucleic acid unidentifiable by the system [58]. Following complete target recognition, conformational changes induced by the binding event serve to lock the surveillance complex onto the DNA and recruit the Cas3 nuclease to degrade the bound target [36, 40, 57, 59].

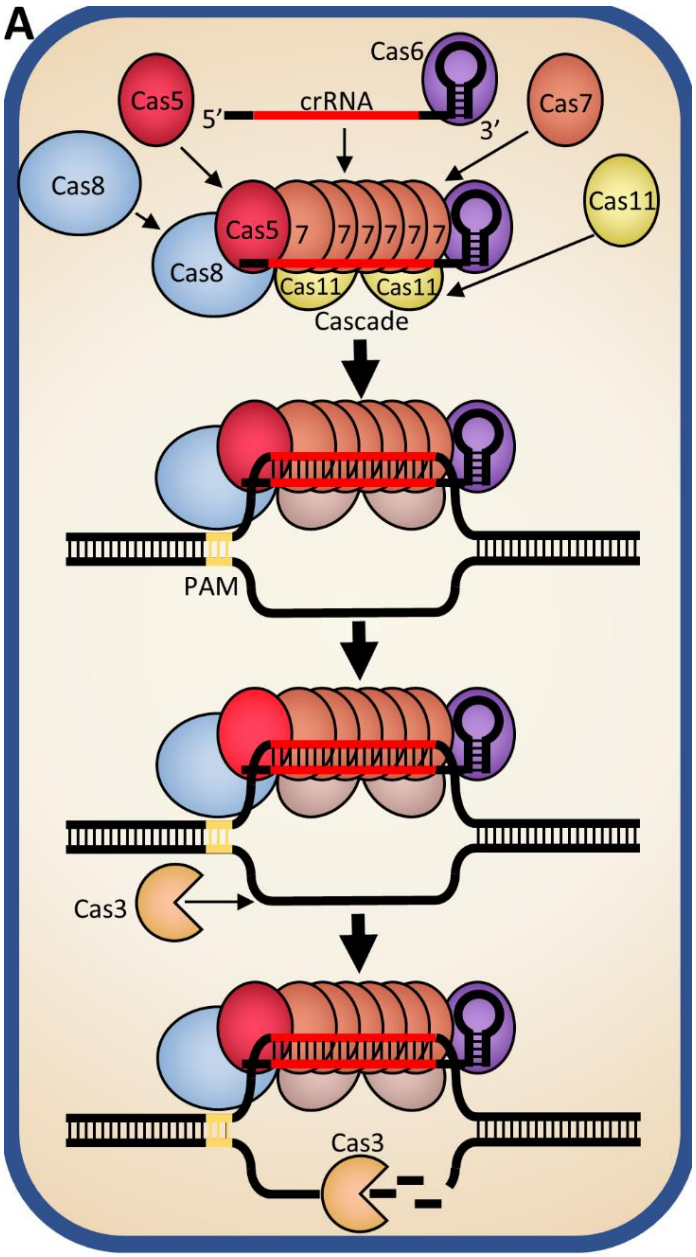


Figure 4. Interference in Type I CRISPR-Cas systems. The Type I surveillance complex (Cascade) is composed of a single copy of Cas5, Cas6, the unique Cas8, multiple Cas7 and Cas11 subunits, and crRNA. The complex surveys dsDNA for a PAM sequence which promotes local melting of DNA for interrogation of the adjacent protospacer. Upon target recognition by complimentary base pairing with the crRNA, Cas3 is recruited to degrade the bound DNA. Figure originally generated by the author and used with permission from Elsevier [39].

Though composition of the Type III surveillance complexes (Figure 5), termed Csm for Type III-A and Cmr for Type III-B, is similar to those of the Type I aCascade complex, their method for targeting invading nucleic acids is significant altered [60-63]. Type III surveillance complexes target both DNA and RNA for degradation by interrogation of mRNA in the transcription bubble of actively transcribed DNA [64-68]. Cleavage of the bound mRNA is performed by the complex's Cas7 backbone while the Cas10 subunit, unique to Type III systems, is responsible for cleavage of the parent DNA [60-62, 68, 69]. Importantly, the Type III interference process is generally performed in a PAM-independent manner, allowing the Type III interference complex to target invading nucleic acids that have undergone a PAM or protospacer mutation that may render them non-targetable by PAM-dependent Type I complexes.

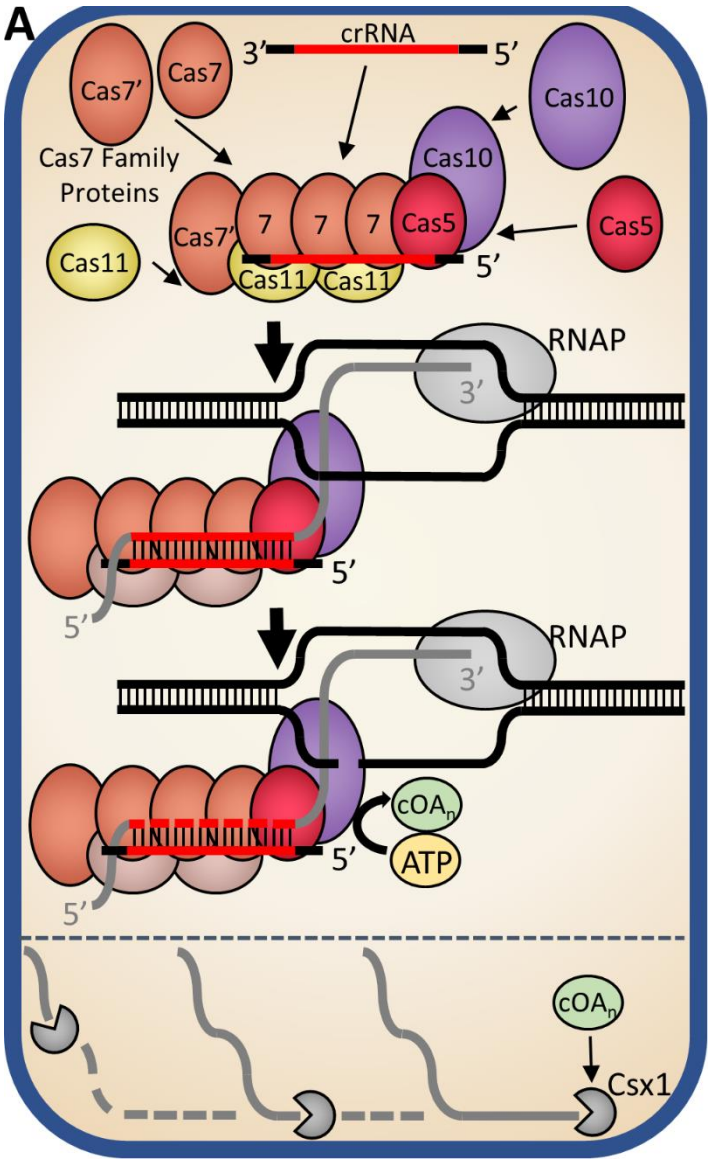


Figure 5. Interference in Type III CRISPR-Cas systems. The Type III surveillance complex (Csm/Cmr) is composed of a single copy of Cas5, the unique Cas10, multiple Cas7 and Cas11 subunits, and crRNA, The complex surveys mRNA in the transcription bubble of actively transcribed DNA. Upon complimentary base pairing with the crRNA, the complex cleaves both the bound mRNA and parent DNA. Figure originally generated by the author and used with permission from Elsevier [39].

Ancillary CRISPR-Cas Proteins and CARF Domains

While great effort has been placed on the study of cas genes directly involved in the active mechanisms of the three stages of the immune response, CRISPR-systems often encode ancillary proteins that have received equivalent attention. Among these are Csa3, associated with Type I-A systems, and Csm6, Csx1, and Csx3 which are often encoded along with Type III systems [13, 70-72]. Early bioinformatics suggested the former of these is transcription factor [70], a hypothesis confirmed by structural analysis performed by Lintner et al. [52]. Their investigations revealed Csa3, encoded by the SSO1445 gene of *S. solfataricus*, harbored an N-terminal modified dinucleotide binding and a C-terminal MarR-like winged helix-turn-helix (wHTH) indicative of DNA binding activity [52]. Two conserved amino acid motifs within the N-terminal dimerization domain, labeled Motif 1 [Thr-h-Gly-Phe-(Asn/Asp)-Glu-X_4-Arg] and Motif 2 [Leu-X_2-Gly-h-Arg], were identified in these structural studies. Interestingly, these motifs span a 2-fold symmetric pocket within the N-terminal modified dinucleotide binding domain, suggesting this this pocket harboring the conserved motifs would be the binding site for a regulatory ligand or second messenger that would allosterically modulate the DNA binding activity of the distal wHTH domains [52]. The Gly-X-Gly-X-X-Ala/Gly motif traditionally conserved in dinucleotide binding domains is absent from Csa3, and superpositional docking of the common ligands bound by dinucleotide binding domains, NAD(P)H and $FADH_2$, revealed significant steric clashes [52]. As such,

Lintner et al. suggested the 2-fold symmetry of this putative ligand binding site in addition to its surface presentation and solvent exposure, and hydrophobic and positively charged moieties, would make the site likely to bind ligands similar to cyclic dinucleotides [52].

Lintner et al. observed that additional ancillary CRISPR-Cas proteins Csx1/Csm6 harbored similar 3D folds, quaternary structure and conserved amino acid motifs to that of the N-terminal dimerization domain of Csa3, suggesting all share a common regulatory binding site [52]. Following these structural investigations, bioinformatics analysis named the modified dinucleotide binding domain shared by the ancillary Csa3 and Cxs1/Csm6 proteins as a CRISPR Associated Rossmann Fold, or CARF domain [73, 74]. Unlike the C-terminal wHTH of Csa3, bioinformatic, structural, and functional investigations of Csx1 and Csm6 revealed C-terminal metal independent HEPN ribonuclease domains [75-77], suggesting regulatory signal or second messenger bound by the CARF domain would activate the RNase activity of Csx1/Csm6 to augment target degradation during the immune response, or to induce cell dormancy and/or programmed cell death [73].

Unique CRISPR-Cas Second Messengers: Cyclic Oligo-Adenylates

Following earlier structural studies that identified CARF family proteins, functional studies of Csx3, a fourth ancillary CRISPR-Cas protein not originally considered a member of the CARF family, identified it as a metal dependent

exoribonuclease [78]. Structural analysis, alongside the functional studies originally labeled Csx3 as a ferredoxin-like fold, and, importantly, revealed a cocrystal structure of Csx3 in complex with a 4-base RNA fragment distal to the exoribonuclease active site [78]. Topuzlu et al. recognized Csx3 as a distant member of the CARF domain superfamily, and connected the earlier hypotheses regarding the putative CARF binding pocket to the 4-base RNA present in the Csx3 cocrystal structure [79]. They suggested the regulatory ligand for CARF domain proteins was likely a symmetric or pseudosymmetric cyclic 4-base RNA [79].

At the same time, structural analysis revealed the Cas10 subunit unique to Type III surveillance complexes contains a cyclase domain [80]. Expanding on this and identifying the missing link in these hypotheses, two simultaneous studies demonstrated that the Cas10 cyclase domain synthesizes cyclic oligoadenylates from ATP when the Type III interference complex binds and identifies a target RNA [81, 82]. During target recognition, Cas10 synthesizes these second messengers in variable lengths, including cyclic tri-adenylate (cA_3), cyclic tetra-adenylate (cA_4), and cyclic hexa-adenylate (cA_6) [81, 82]. Additionally, these groups demonstrated that these cyclic oligoadenylate signals modulate the endonuclease activities of the HEPN domains of Csx1 and Csm6 [81, 82], confirming earlier hypotheses regarding Csx1, Csm6, and likely all CARF domain proteins [52, 73, 79]. The ability for Cas10 to synthesize varied cyclic oligoadenylate lengths was identified as a key mechanism as different lengths were required to stimulate ribonuclease activity of Csm6 from different organisms. Csm6 from S. *thermophilus* required cA_6 for

activation while those from *T. thermophilus and E. italicus* required cA_4 [81, 82]. Halting cA_4-stimulated activation of nuclease activity through mutation to the CARF binding pocket in EiCsm6 confirmed the hypothesis that the conserved binding pocket harbored there was responsible for second cA_4 binding and allosteric control over the HEPN effector domain [82].

The role of Csa3a in CRISPR-Cas

After the initial structural studies of Csa3 by Lintner et al., the family was divided into the Csa3a and Csa3b clades [83, 84]. Transcriptomic analysis by Liu et al. revealed overexpression of Csa3a in *S. islandicus* Rey15A increases transcript levels for acquisition genes *cas1*, *cas2*, *cas4a* and *csa1*, and results in acquisition of new spacers [83, 85] outside of viral infection. DNA foot-printing identified Csa3a binding sites immediately upstream of the *cas4a/cas1/cas2* acquisition cassette and several CRISPR-loci [85], suggesting Csa3a regulates expression of CRISPR loci and cas genes involved in spacer acquisition. *S. solfataricus* Sso1445, the protein of interest in our work as well as of the initial structural studies performed by Lintner et al., was identified as a member of the Csa3a subfamily [52, 83].

Csa3a and cA_4: a Nexus for Class 1 Systems

The sum of works regarding CARF domain proteins, their second messenger in the form of the Cas10-synthesized cyclic oligoadenylates, and the Type I associated Csa3a transcription factor suggests an interaction between Csa3a and cA_4 would represent a nexus for the integration of Type I and Type III systems into

a coordinated Class1 response in organisms that harbor both types [28]. Synthesis of cA_4 by the Type III interference complexes may stimulate primed acquisition of new spacers by activation of Csa3a, in turn upregulating transcription of the acquisition cassette containing cas1, cas2, and cas4 as well as expression of newly acquired spacers through upregulation of CRISPR loci transcription [28, 52, 85].

In this work we confirm the presence of this nexus connecting the Type I and Type III systems in *S. solfataricus* by showing SsCsa3a does indeed recognize cA_4. Structural investigations reveal conserve motifs are responsible for cA_4 recognition and this interaction results in significant cA_4 induced conformational changes in Csa3a. Further, we show Csa3a lacks the self-regulatory ring nuclease activity exhibited by other CARF proteins and characterize DNA recognition in the absence and presence of cA_4. Overall, the role of cA_4 and Csa3a in the regulation of CRISPR-Cas may be more complicated than previously appreciated.

High Level Protein Expression For the Study of Human Galectins and Undergraduate Research Pursuits

Structural studies of CRISPR-cas regulation resulted in development of a workflow for the generation of large amounts of pure protein, as required for the variety of assays and crystal growth. The requirement of significant quantities of protein is not a problem unique to CRISPR-cas or crystallography, but is faced in most research endeavors pertaining to the characterization of proteins, their activities, or their place in a biological system. Projects involving commonly studied

proteins often rely on commercially expressed and purified constructs that impose significant costs and restrict researcher control over protein quality, quantity, available mutants/constructs, and many other characteristics. The ability to produce necessary quantities of protein in a laboratory setting opens avenues of investigation typically closed to laboratories or researchers reliant on commercially produced proteins. The same principles and workflow used to generate large quantities of protein for CRISPR-Cas studies was applied to the generation of *Homo sapiens* Galectins 3, 7, and 8 along with multiple shortened Galectin-3 constructs.

While this project initially began not as a research project of its own, but simply as an endeavor for generating large quantities of pure Galectins, we observed a near order-order-of magnitude yield increase of protein using our autoinduction method compared to the traditional IPTG-induction method, thus, we designed experiments to pursue the mechanisms behind this phenomenon. Our investigations led to the identification of a putative inhibitory feedback mechanism that could potentially impact a wide selection of carbohydrate-binding proteins.

In addition to its use for protein stock generation, this project was used to test the efficacy of large-scale curriculum-based undergraduate research (CURE). As numbers of biochemistry undergraduate students grow, the need for research opportunities does as well. Traditionally, one or more undergraduates will work in a research lab at a time, often supervised by a graduate student. While this is sufficient for meeting the research requirement for a single student, it can be

difficult to accommodate large numbers of undergraduate researchers and progress on a project is limited by a student's schedule.

The workflow used to generate Galectin proteins for use in biochemical assays was altered to be approachable by undergraduate researchers who work in teams of two to five students to progress through an expression and purification throughout a week without the usual time lapses introduced by a single student's schedule. After initial instruction by a graduate student or professor, students in this setting were able to teach the parts of the process they were responsible for to other students, reducing the long-term workload on primary researchers. In this manner, up to twelve students at a time developed research skills while making progress on a research project equivalent to multiple full-time researchers.

References

1. Brussaard, C.P.D., A.-C. Baudoux, and F. Rodríguez-Valera, *Marine Viruses*, in *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential*, L.J. Stal and M.S. Cretoiu, Editors. 2016, Springer International Publishing: Cham. p. 155-183.

2. Enikeeva, F.N., K.V. Severinov, and M.S. Gelfand, *Restriction–modification systems and bacteriophage invasion: Who wins?* Journal of Theoretical Biology, 2010. **266**(4): p. 550-559.

3. Xu, T., et al., *A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella.* Nucleic Acids Research, 2010. **38**(20): p. 7133-7141.

4. Labrie, S.J., J.E. Samson, and S. Moineau, *Bacteriophage resistance mechanisms.* Nature Reviews Microbiology, 2010. **8**(5): p. 317-327.

5. Nordström, K. and A. Forsgren, *Effect of protein A on adsorption of bacteriophages to Staphylococcus aureus.* J Virol, 1974. **14**(2): p. 198-202.

6. Hammad, A.M.M., *Evaluation of alginate-encapsulated Azotobacter chroococcum as a phage-resistant and an effective inoculum*. Journal of Basic Microbiology, 1998. **38**(1): p. 9-16.

7. Destoumieux-Garzón, D., et al., *The iron–siderophore transporter FhuA is the receptor for the antimicrobial peptide microcin J25: role of the microcin Val11–Pro16 β-hairpin region in the recognition mechanism*. Biochemical Journal, 2005. **389**(3): p. 869-876.

8. Hille, F., et al., *The Biology of CRISPR-Cas: Backward and Forward*. Cell, 2018. **172**(6): p. 1239-1259.

9. Sorek, R., C.M. Lawrence, and B. Wiedenheft, *CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea*. Annual Review of Biochemistry, 2013. **82**(1): p. 237-266.

10. Barrangou, R. and P. Horvath, *A decade of discovery: CRISPR functions and applications*. Nat Microbiol, 2017. **2**: p. 17092.

11. Jackson, S.A., et al., *CRISPR-Cas: Adapting to change*. Science, 2017. **356**(6333).

12. Koonin, E.V. and K.S. Makarova, *Origins and evolution of CRISPR-Cas systems*. Philos Trans R Soc Lond B Biol Sci, 2019. **374**(1772): p. 20180087.

13. 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.

14. McGinn, J. and L.A. Marraffini, *Molecular mechanisms of CRISPR-Cas spacer acquisition*. Nat Rev Microbiol, 2019. **17**(1): p. 7-12.

15. Mohanraju, P., et al., *Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems*. Science, 2016. **353**(6299): p. aad5147.

16. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes*. Science (New York, N.Y.), 2007. **315**(5819): p. 1709-12.

17. Bolotin, A., et al., *Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin*. Microbiology, 2005. **151**(Pt 8): p. 2551-61.

18. Groenen, P.M., et al., *Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method*. Mol Microbiol, 1993. **10**(5): p. 1057-65.

19. Ishino, Y., et al., *Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.* J Bacteriol, 1987. **169**(12): p. 5429-33.

20. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes.* Mol Microbiol, 2002. **43**(6): p. 1565-75.

21. Makarova, K.S., et al., *A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.* Biology direct, 2006. **1**: p. 7.

22. Marraffini, L.A. and E.J. Sontheimer, *CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.* Science (New York, N.Y.), 2008. **322**(5909): p. 1843-5.

23. Mojica, F.J., et al., *Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.* J Mol Evol, 2005. **60**(2): p. 174-82.

24. 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.

25. 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, 2005. **151**(Pt 3): p. 653-63.

26. Sorek, R., C.M. Lawrence, and B. Wiedenheft, *CRISPR-mediated adaptive immune systems in bacteria and archaea.* Annu Rev Biochem, 2013. **82**: p. 237-66.

27. Hille, F., et al., *The Biology of CRISPR-Cas: Backward and Forward.* Cell, 2018. **172**(6): p. 1239-1259.

28. Charbonneau, A.A., C.C. Gauvin, and C.M. Lawrence, *CRISPR-Cas, A Prokaryotic Adaptive Immune System*, in *Encyclopedia of Biological Chemistry, 3rd Edition*, J. Joseph, Editor. 2021, Oxford: Elsevier. p. 717-741.

29. Anders, C., et al., *Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease.* Nature, 2014. **513**(7519): p. 569-73.

30. Jiang, W., et al., *Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids.* PLOS Genetics, 2013. **9**(9): p. e1003844.

31. Stern, A., et al., *Self-targeting by CRISPR: gene regulation or autoimmunity?* Trends in Genetics, 2010. **26**(8): p. 335-340.

32. Agari, Y., et al., *Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection.* Journal of Molecular Biology, 2010. **395**(2): p. 270-281.

33. Fusco, S., et al., *Transcriptome analysis of Sulfolobus solfataricus infected with two related fuselloviruses reveals novel insights into the regulation of CRISPR-Cas system.* Biochimie, 2015. **118**: p. 322-332.

34. Perez-Rodriguez, R., et al., *Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli.* Molecular Microbiology, 2011. **79**(3): p. 584-599.

35. Young, J.C., et al., *Phage-Induced Expression of CRISPR-Associated Proteins Is Revealed by Shotgun Proteomics in Streptococcus thermophilus.* PLOS ONE, 2012. **7**(5): p. e38077.

36. Jore, M.M., et al., *Structural basis for CRISPR RNA-guided DNA recognition by Cascade.* Nat Struct Mol Biol, 2011. **In Press**.

37. 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).* Journal of Biological Chemistry, 2011. **286**(24): p. 21643-21656.

38. Sashital, D.G., M. Jinek, and J.A. Doudna, *An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3.* Nat Struct Mol Biol, 2011. **18**(6): p. 680-7.

39. Charbonneau, A.A., C.C. Gauvin, and C.M. Lawrence, *Methods | CRISPR-Cas, A Prokaryotic Adaptive Immune System*, in *Encyclopedia of Biological Chemistry III (Third Edition)*, J. Jez, Editor. 2021, Elsevier: Oxford. p. 717-741.

40. Brouns, S.J., et al., *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science (New York, N.Y.), 2008. **321**(5891): p. 960.

41. Deltcheva, E., et al., *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.* Nature, 2011. **471**(7340): p. 602-7.

42. Hale, C.R., et al., *RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex.* Cell, 2009. **139**(5): p. 945-56.

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

44. Sapranauskas, R., et al., *The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli.* Nucleic Acids Res, 2011. **39**(21): p. 9275-82.

45. Sakai, H.D. and N. Kurosawa, *Saccharolobus caldissimus gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of Sulfolobus solfataricus as Saccharolobus solfataricus comb. nov. and Sulfolobus shibatae as Saccharolobus shibatae comb. nov.* Int J Syst Evol Microbiol, 2018. **68**(4): p. 1271-1278.

46. Charbonneau, A.A., et al., *Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response in Saccharolobus solfataricus.* Biomolecules, 2021. **11**(12): p. 1852.

47. Kraft, P., et al., *Structure of D-63 from Sulfolobus Spindle-Shaped Virus 1: Surface Properties of the Dimeric Four-Helix Bundle Suggest an Adaptor Protein Function.* Journal of Virology, 2004. **78**(14): p. 7438-7442.

48. Larson, E.T., et al., *Structure of A197 from Sulfolobus Turreted Icosahedral Virus: a Crenarchaeal Viral Glycosyltransferase Exhibiting the GT-A Fold.* Journal of Virology, 2006. **80**(15): p. 7636-7644.

49. Lintner, N.G., et al., *The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System.* Journal of Molecular Biology, 2011. **405**(4): p. 939-955.

50. 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.

51. Menon, S.K., et al., *The crystal structure of D212 from sulfolobus spindle-shaped virus ragged hills reveals a new member of the PD-(D/E)XK nuclease superfamily.* J Virol, 2010. **84**(12): p. 5890-7.

52. Lintner, N.G., et al., *The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System.* J Mol Biol, 2011. **405**(4): p. 939-955.

53. Lillestøl, R.K., et al., *CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties.* Molecular Microbiology, 2009. **72**(1): p. 259-272.

54. Zink, I.A., E. Wimmer, and C. Schleper, *Heavily Armed Ancestors: CRISPR Immunity and Applications in Archaea with a Comparative Analysis of CRISPR Types in Sulfolobales.* Biomolecules, 2020. **10**(11).

55. Hayes, R.P., et al., *Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli.* Nature, 2016. **530**(7591): p. 499-503.

56. Xiao, Y., et al., *Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System.* Cell, 2017. **170**(1): p. 48-60 e11.

57. Jackson, R.N., et al., *Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli.* Science, 2014. **345**(6203): p. 1473-9.

58. Semenova, E., et al., *Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.* Proc Natl Acad Sci U S A, 2011. **108**(25): p. 10098-103.

59. Mulepati, S., A. Heroux, and S. Bailey, *Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target.* Science, 2014. **345**(6203): p. 1479-84.

60. Osawa, T., et al., *Crystal Structure of the CRISPR-Cas RNA Silencing Cmr Complex Bound to a Target Analog.* Mol Cell, 2015. **58**(3): p. 418-30.

61. Staals, R.H., et al., *RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus.* Mol Cell, 2014. **56**(4): p. 518-30.

62. Taylor, D.W., et al., *Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning.* Science, 2015. **348**(6234): p. 581-5.

63. Zhao, H., et al., *Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli.* Nature, 2014. **515**(7525): p. 147-50.

64. Deng, L., et al., *A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus.* Mol Microbiol, 2013. **87**(5): p. 1088-99.

65. Elmore, J.R., et al., *Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system.* Genes Dev, 2016. **30**(4): p. 447-59.

66. Estrella, M.A., F.T. Kuo, and S. Bailey, *RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex.* Genes Dev, 2016. **30**(4): p. 460-70.

67. Kazlauskiene, M., et al., *Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition.* Mol Cell, 2016. **62**(2): p. 295-306.

68. Samai, P., et al., *Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity.* Cell, 2015. **161**(5): p. 1164-1174.

69. Tamulaitis, G., et al., *Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus.* Mol Cell, 2014. **56**(4): p. 506-17.

70. Haft, D., et al., *A Guild of 45 CRISPR-Associated (Cas) Protien Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.* Plos Computational Biology, 2005. **1**(6): p. 474-483.

71. Makarova, K.S. and E.V. Koonin, *Annotation and Classification of CRISPR-Cas Systems.* Methods Mol Biol, 2015. **1311**: p. 47-75.

72. Athukoralage, J.S., et al., *Tetramerisation of the CRISPR ring nuclease Crn3/Csx3 facilitates cyclic oligoadenylate cleavage.* Elife, 2020. **9**.

73. Makarova, K.S., et al., *CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems.* Front Genet, 2014. **5**: p. 102.

74. Makarova, K.S., et al., *Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense.* Nucleic Acids Res, 2020. **48**(16): p. 8828-8847.

75. Anantharaman, V., et al., *Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing.* Biol Direct, 2013. **8**: p. 15.

76. Kim, Y.K., Y.G. Kim, and B.H. Oh, *Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of Pyrococcus furiosus.* Proteins, 2013. **81**(2): p. 261-70.

77. Sheppard, N.F., et al., *The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease.* RNA, 2016. **22**(2): p. 216-24.

78. Yan, X., W. Guo, and Y.A. Yuan, *Crystal structures of CRISPR-associated Csx3 reveal a manganese-dependent deadenylation exoribonuclease.* RNA Biol, 2015. **12**(7): p. 749-60.

79. Topuzlu, E. and C.M. Lawrence, *Recognition of a pseudo-symmetric RNA tetranucleotide by Csx3, a new member of the CRISPR associated Rossmann fold superfamily.* RNA Biol, 2016. **13**(2): p. 254-7.

80. Zhu, X. and K. Ye, *Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems.* FEBS Lett, 2012. **586**(6): p. 939-45.

81. Kazlauskiene, M., et al., *A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems.* Science, 2017. **357**(6351): p. 605-609.

82. Niewoehner, O., et al., *Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers.* Nature, 2017. **548**(7669): p. 543-548.

83. Liu, T., et al., *Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition.* Nucleic Acids Res, 2015. **43**(2): p. 1044-55.

84. He, F., et al., *CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b.* Nucleic Acids Res, 2017. **45**(4): p. 1902-1913.

85. Liu, T., et al., *Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus.* Nucleic Acids Res, 2017. **45**(15): p. 8978-8992.

CHAPTER TWO

METHODS AND METHODOLOGY

Expanded Methods for Manuscript 1

Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response in *Saccharolobus solfataricus*

The methods described in this chapter are expanded procedures or information not found in the original manuscript.

Bioinformatics

Identification of conserved residues unique to Csa3a were identified by alignment of annotated Csa3a sequences in MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [1]. Notably, this alignment did not include Csa3b sequences like previously published alignments.[2]

Sulfolobus Strains, Growth Conditions

S. solfataricus PH1-16 (Δ*pyrEF*) was cultured in DT media (with 20 µg/mL uracil) at 78 °C and shaking at 65 rpm. An empty pSeSD1 vector and pSeSD1 containing *S. solfataricus* P2 csa3a (pSeSD-sso1445) were transformed into PH1-16 cells via electroporation using a BioRad GenePulser Xcell [3, 4]. 50 mL *S. solfataricus* PH1-16 cells for each transformation were grown to an OD_{650} ~0.8, pelleted via centrifugation at 3000 xg for 10 minutes, and gently resuspended in 50

mL ice cold, filter sterilized 20 mM sucrose. Cells were pelleted again via centrifugation and resuspended in 25 mL, 10 mL, 1 mL, and 0.4 mL ice cold, filter sterilized 20 mM sucrose to remove any salts that might cause arcing during electroporation. ~1 µg was dialyzed on a 0.2 µm filter floated on a deionized H_2O for 30 min. The vector and 50 µL of washed cells were pipetted into a sterile, cold 1 mm gap electroporation cuvette and incubated on ice for 10 min. Cells were electroporated at 1.5 kV, 400 Ω, 25 kF, resulting in time constants of 9 – 10 ms. Transformed cells were immediately transferred to 1 mL either 78 °C or ice cold DT media without uracil and incubated at 78 °C for 1 hour before being used to inoculate 50 mL starter cultures without uracil of each transformant. Transformants were selected by monitoring uracil auxotrophy in DT media.. Cultured transformants were plated on 0.7% GelRite DT media plates and incubated at 78°C until colony formation. A monoculture of transformed cells was started via inoculation of DT media with a single colony **[4]**. Glycerol stocks of transformed monocultures were created by mixing 500 µL log-phase (OD_{650} = 0.3-0.5) transformed *S. solfataricus* PH1-16 culture with 500 µL 100% glycerol. Glycerol stocks were stored at -80 °C.

Csa3a Expression and Purification in Sulfolobus Solfataricus

Methods for protein expression in *S. solfataricus* were adapted from F. He et al. [5]. A starter culture was started by inoculating 50 mL DT media with a single colony of *S. solfataricus* PH1-16 transformed with the pSeSD-sso1445 vector.

Starter culture growth was monitored, along with a negative control to ensure uracil auxotrophy was abolished by the presence of the vector. Vector presence was also verified via PCR. 5 mL of the starter culture grown to OD_{650} 0.6 was used to inoculate a 250 mL culture which was grown until OD_{650} reached 0.6 – 0.8. Protein expression was induced by addition of 0.5 mL filter sterilized 10% Arabinose (final concentration 0.2%). Cultures were incubated at 78 °C with shaking at 65 rpm for 48 hr. Cells were harvested by centrifugation at 5500 x g for 20 min, and the cell pellets were stored at −80 °C. Protein purification was performed as for *E. coli*.

Screening for Csa3a-cA_4 Crystallization Conditions

To identify conditions for crystallization of the Csa3-cA_4 complex, purified Csa3 was concentrated to 15 mg/mL with 10,000 MWCO Amicon Ultra™ spin concentrators (Millipore) for the initial conditions tested – those used by Lintner et. al [2] to crystallize the Csa3 in the absence of cA_4 (2 uL 15 mg/mL Csa3 and 2uL 37.5-42.5% MPD, 5% PEG-8000, and 0.1 M sodium cacodylate at pH 7.0-7.5) with an additional 0.2 uL 10 mM cA_4 (c-tetraAMP, BioLog Lifescience Institute) in hanging drop vapor diffusion at 22°C. Addition of cA_4 under these conditions resulted in rapid precipitation of protein.

Crystal soaks were attempted using crystals of the protein grown under the conditions used by Lintner et. al [2]. 0.2 uL of 10 mM cA_4 or A_4>P (ChemGenes Corporation) was added to drops containing fully grown crystals. Crystals soaked with this method degraded rapidly in the drop.

Crystal screens were set up using a TTPLabtech Mosquito® Crystal crystallization robot. Conditions were screened by sitting drop vapor diffusion consisting of 100 nL Csa3 at 7.5 mg/mL, 50 nL 2mM cA_4, and 200 nL of well solution from a variety of crystallization screens. Crystallization screens included Hampton Crystal Screen 1, Hampton Crystal Screen 2, Hampton PEG Ion 1, Hampton PEG Ion 2, Hampton Salt Rx 1, Hampton Salt Rx 2, Hampton Crystal Screen Lite, Hampton Crystal Screen Cryo, Hampton Peg Rx 1, Hampton Peg Rx2, Hampton Index, and Emerald BioStructures Cryo. Control screens were set up using the same conditions for protein and well solution but replacing the cA_4 with 50 nL water. Drops were incubated at 22 °C for 1-2 days and multiple conditions yielded crystals and microcrystals, though few were successfully scaled up to 4 uL hanging drop vapor diffusion to generate larger crystals for data collection.

Quantitation of EMSA Binding Assays

To acquire quantitative binding data, gel images were analyzed using ImageJ image analysis software. Unshifted bands were isolated as individual "lanes" on the gel image the total pixel intensity for each was recorded using the "Plot Lanes" tool to count peak intensities using identical peak widths (Figure 1). To calculate the percent of the DNA bound for each band the total intensity of the band was subtracted from the total intensity of the 0 nM Csa3 band and the resulting intensity was divided by the total intensity of the 0 nM Csa3 band. The calculations for percent DNA bound assume that the total intensity of the 0 nM band represents

100% of the DNA loaded into the well. Using GraphPad Prism, a non-linear curve was fit to the calculated percent bound-DNA using the "Specific Binding with Hill Slope" function: $Y=B_{max}*X^h/(K_d^h + X^h)$ where Y is the percent bound-DNA, X is the corresponding Csa3 concentration in nM, B_{max} is the maximum specific binding in percent (constrained to 100%), K_d is the concentration of Csa3 in nM needed to achieve half-maximum binding, and h is the Hill coefficient.

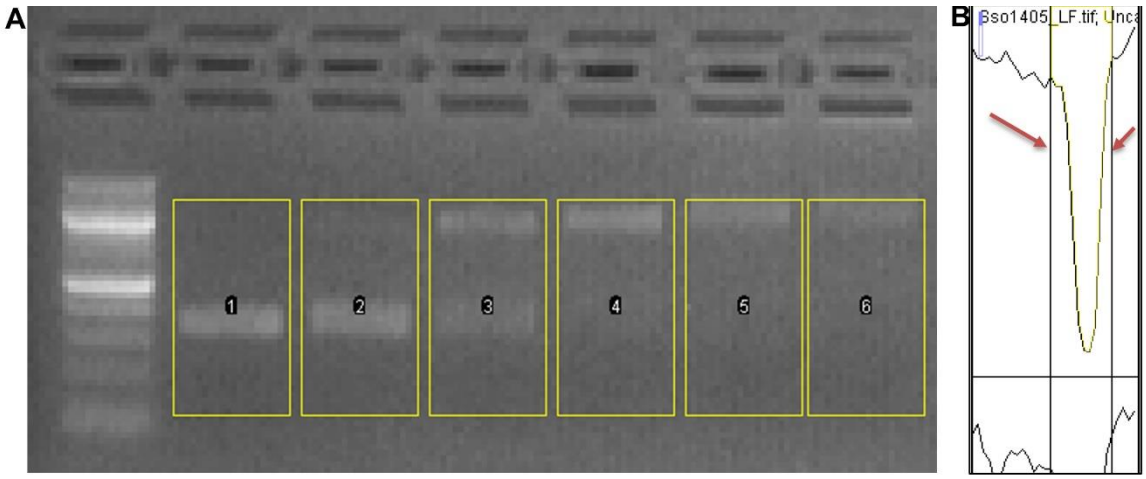


Figure 1 ImageJ Analysis of EMSA Gels. A. 32-bit image of the EMSA gel with yellow boxes indicating the "lane" selection areas for analysis. Only the lower, unshifted band is necessary for quantitation. B. The "Plot Lanes" interface shows peaks for pixel intensity across the selected lanes (only lane 1 is fully visible in the figure). Red arrows indicate lines drawn perpendicular to the lane boundary in order to isolate the single peak that represents the unshifted band. The yellow-highlighted area indicates the area selected with the "wand" tool to calculate total intensity for the peak.

References

1. Madeira, F., et al., *Search and sequence analysis tools services from EMBL-EBI in 2022.* Nucleic acids research, 2022. **50**(W1): p. W276-W279.

2. Lintner, N.G., et al., *The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System.* J Mol Biol, 2011. **405**(4): p. 939-955.

3. Peng, N., et al., *A synthetic arabinose-inducible promoter confers high levels of recombinant protein expression in hyperthermophilic archaeon Sulfolobus islandicus.* Appl Environ Microbiol, 2012. **78**(16): p. 5630-7.

4. 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.

5. He, F., et al., *CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b.* Nucleic Acids Res, 2017. **45**(4): p. 1902-1913.

CHAPTER THREE

CYCLIC TETRA-ADENYLATE (CA_4) RECOGNITION BY CSA3; IMPLICATIONS FOR AN INTEGRATED CLASS 1 CRISPR-CAS IMMUNE RESPONSE IN *S. SOLFATARICUS*

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Alexander A. Charbonneau

Contributions: Conceptualization, methodology, formal analysis, investigation, writing – original draft preparation, writing – review and editing, visualization.

Co-Author: Debra M. Eckert

Contributions: Conceptualization, methodology, formal analysis, investigation, writing – original draft preparation, writing – review and editing, visualization.

Co-Author: Colin C. Gauvin

Contributions: Conceptualization, methodology, formal analysis, investigation, writing – original draft preparation, writing – review and editing, visualization.

Co-Author: Nathaniel G. Lintner

Contributions: Methodology.

Co-Author: C. Martin Lawrence

Contributions: Conceptualization, methodology, writing – original draft preparation, writing – review and editing, project administration, funding acquisition.

Manuscript Information

Alexander A. Charbonneau[1,2], Debra M. Eckert[3‡], Colin C. Gauvin[1,2‡], Nathanael G. Lintner[§], and C. Martin Lawrence[1,2,*]

[1] Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717 USA
[2] Thermal Biology Institute, Montana State University, Bozeman, MT 59717 USA
[3] University of Utah School of Medicine, Salt Lake City, UT, USA
[§] Current address: Medicine Design, Pfizer, Cambridge, MA 02139
[‡] These authors contributed equally.
***** Correspondence: lawrence@montana.edu; Tel.: +1 406-994-5382

Biomolecules

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

MDPI
9 December 2021
Biomolecules 2021, Volume 11
https://doi.org/10.3390/biom11121852

Abstract

Csa3 family transcription factors are ancillary CRISPR associated proteins composed of N-terminal CARF domains and C-terminal winged helix-turn-helix domains. The activity of Csa3 transcription factors is thought to be controlled by cyclic oligoadenyate(cOA) second messengers produced by type III CRISPR-Cas surveillance complexes. Here we show that *S. solfataricus* Csa3a recognizes cyclic tetra-adenylate (cA_4) and that Csa3a lacks self-regulating "ring nuclease" activity present in some other CARF domain proteins. The crystal structure of the Csa3a/cA4 complex was also determined and the structural and thermodynamic basis for cA_4 recognition are described, as are conformational changes in Csa3a associated with cA_4 binding. We also characterized the effect of cA_4 on recognition of putative DNA binding sites. Csa3a binds to putative promoter sequences in a nonspecific, cooperative and cA_4 independent manner, suggesting a more complex mode of transcriptional regulation. We conclude the Csa3a/cA_4 interaction represents a nexus between the type I and type III CRISPR-Cas systems present in *S. solfataricus*, and discuss the role of the Csa3/cA_4 interaction in coordinating different arms of this integrated class 1 immune system to mount a synergistic, highly orchestrated immune response.

1. Introduction

The basic tenets of CRISPR-Cas adaptive immunity are now well established [1-3]. This heritable, adaptive immune response is manifested in 3 distinct stages, i)

adaptation, ii) crRNA maturation and iii) interference. During adaptation, the CRISPR associated proteins Cas1, Cas2 and Cas4 integrate short segments of invading DNA (protospacers) into CRISPR loci. These "spacers" are separated from those of previous invaders by a short DNA repeat, giving rise to the namesake Clustered Regularly Interspaced Short Palindromic Repeats, or CRISPR loci. In stage ii) the CRISPR-locus is transcribed and the pre-CRISPR RNA is processed by Cas proteins (Cas6 or Cas12) or cellular nucleases to produce CRISPR-RNA (crRNA). In stage iii) crRNA and Cas proteins assemble into nucleoprotein complexes that survey the cell for invading nucleic acid. Upon recognition, Cas nuclease activity then degrades the invading genetic material.

Also well appreciated is the extreme diversity of CRISPR-Cas systems. They are grouped into 2 distinct classes, Class 1 and Class 2, and further subdivided into 6 types and 24 different subtypes [4]. Relevant to the work described here are subtypes I-A, III-B, and III-D belonging to class 1 systems. The *Saccharolobus solfataricus* P2 genome (formerly *Sulfolobus solfataricus* [5]) encodes three Type I-A, three Type III-B, and one Type III-D module, and harbors six CRISPR loci designated CRISPRs A–F (Supplemental Figure S1) [6-8]. Type I-A effector complexes, also known as archaeal Cascade (aCascade), target double stranded DNA in a PAM dependent manner [1, 9], while the type III-B and III-D effector complexes target transcriptionally active DNA and mRNA in a PAM independent manner.

1.1 CARF Domains

CRISPR systems frequently include ancillary components. These include Csa3 which is associated with Type I-A systems, and Csm6, Csx1 and Csx3 which are associated with type III systems [4, 10-12]. Consistent with early bioinformatics, structural analysis of Csa3 revealed a modified dinucleotide binding domain at the N-terminus, and a MarR-like winged helix-turn-helix at the C-terminus, indicative of DNA binding [6]. The structural analysis also identified two conserved sequence motifs in the N-terminal dimerization domain [Thr-h-Gly-Phe-(Asn/Asp)-Glu-X_4-Arg and Leu-X_2-Gly-h-Arg]. These conserved motifs are distal to the DNA binding domains, and form a 2-fold symmetric pocket that spans the dimer interface. This conserved pocket exhibits the hallmarks of a regulatory ligand-binding site, suggesting Csa3 is a transcriptional regulator under allosteric control of the N-terminal domain [6]. It was suggested that the 2-fold symmetric pocket with hydrophobic and positively charged side chains might be complementary to signals similar to those of cyclic dinucleotides [6].

Lintner et al. also noted a similar 3D fold, quaternary structure and conserved sequence motifs at the dimer interface of Csx1/Csm6, a second class of CRISPR/Cas ancillary protein, suggesting a similar regulatory binding site in Csx1/Csm6 [6]. Ensuing bioinformatics work rechristened this modified dinucleotide binding domain in Csa3 and Cxs1/Csm6 as the CRISPR Associated Rossmann Fold, or CARF domain [13, 14]. Bioinformatics, structural and functional work also identified the C-terminal domains of Csx1 and Csm6 as metal independent HEPN

ribonucleases [15-17], suggesting the CARF domain regulatory signal would activate Csx1/Csm6 RNase activity to complement CRISPR-Cas action, or to induce cell dormancy and/or programmed cell death [13].

1.2 Cyclic Oligo-Adenylate Second Messengers

In subsequent work, a fourth CRISPR/Cas ancillary protein known as Csx3 was shown to be a metal dependent exoribonuclease, and structural analysis of the Csx3/RNA cocrystal identified a 4-base RNA fragment distant from the active site [18]. Importantly, Csx3 was then recognized as a distant member of the CARF domain superfamily, and a cyclic 4-base RNA was suggested as the regulatory ligand for CARF domain proteins [19].

In parallel, structural studies had also identified a cyclase domain in Cas10, the large subunit of type III surveillance complexes [20]. In beautiful work, two different groups experimentally connected the dots, demonstrating that the Cas10 cyclase domains utilize ATP to synthesize cyclic tetra-adenylate (cA4) and hexa-adenylate (cA6) when the type III effector complex is bound to target RNA. They also demonstrated these signals do indeed stimulate the endonuclease activities of Csx1 and Csm6 respectively [21, 22], confirming the earlier hypotheses [6, 13] for Csx1 and Csm6.

1.3 Transcriptional Regulation by Csa3

Two Csa3 clades have been identified, 3a and 3b [23, 24]. Liu et al. have shown overexpression of Csa3a in *S. islandicus* Rey15A increases transcript levels

for *cas1*, *cas2*, *cas4a* and *csa1*, resulting in acquisition of new spacers [23, 25]. Further, DNA foot-printing identified Csa3a binding sites immediately upstream of the *cas4a/cas1/cas2* acquisition cassette and several CRISPR-loci [25], suggesting Csa3a regulates expression of CRISPR loci and cas genes involved in spacer acquisition. They also identified *S. solfataricus* Sso1445, the subject of our work, as a member of the Csa3a subfamily [23].

1.4 Integration of Class 1 Systems

Collectively, the work above suggests the cA_4 second message generated by Type-III systems may act through Csa3a to provide an integrated Class 1 response in organisms harboring both Type-1-A and Type-III systems [3]. Specifically, production of the cA_4 second messenger may stimulate primed acquisition of new spacers by upregulating transcription of the acquisition cassette (cas1, cas2 and cas4), and expression of the new spacers by upregulating transcription of the CRISPR loci as well [3, 6, 25]. Here we show that *S. solfataricus* Csa3a does indeed recognize cA_4. We describe the structural basis for cA_4 recognition and cA_4 induced conformational changes in Csa3a, and characterize DNA recognition in the absence and presence of cA_4. Overall, the role of cA_4 and Csa3a in the regulation of CRISPR-Cas may be more complicated than previously appreciated. With regard to genome engineering, we note that greater understanding of transcriptional regulation by cyclic oligoadenylates might enable engineered transcriptional responses upon detection of specific target genes.

2. Materials and Methods

2.1 Bioinformatics

Gene annotation was obtained from NCBI RefSeq when available, or, when no annotation was available, the ORF amino acid translation was acquired through UniproUGENE, and submitted to HHPred for profile-profile sequence alignment. Identification was considered conclusive only when the E-value was <0.001 and probability was greater than 95%. In the event no conclusion could be made, genes were submitted to HHBoost machine learning server for further aid in identification, or left as un-annotated [26].

Putative *S. solfataricus* Csa3a binding sites in the Cas4a (Sso1451) promoter and the leader sequences of CRISPR loci C, D and E were identified by Liu et al, 2017 [25]. Similarly, we also analyzed putative acquisition gene promoters and CRISPR leader sequences, submitting them to the Multiple EM for Motif Elicitation server. The results were trimmed and resubmitted until only conserved bases were remaining. The conserved motif was then queried against all available fully sequenced *Sulfolobales* genomes, and identified locations were annotated locally.

2.2 Csa3a Expression and Purification

Expression and purification of SsCsa3a (Sso1445) was performed in *E. coli* as previously described [6, 27]. Briefly, BL21(DE3)-pRIL *E. coli* (Stratagene) were transformed with the pEXP14-6×His-Sso1445 expression plasmid [6] and a single colony was used to inoculate 5 ml of LB with 100 µg/ml ampicillin and 34 µg/ml

chloramphenicol and grown overnight. 750 ml of ZYP-5052 autoinduction media [28] containing 100 µg/ml ampicillin and 34 µg/ml chloramphenicol was then inoculated with 750 µl of the starter culture and grown at 37 °C for 16–20 h. Cells were harvested by centrifugation at 5500 x g for 20 min, and the pellets were stored at −80 °C.

Cell pellets were thawed and resuspended in 5 ml of lysis buffer (20 mM Tris pH 8.0, 400 mM NaCl, 0.1 mM PMSF) per gram of cell pellet. Cells were lysed by passage through a French press (American Instrument Co., Inc., Silver Springs, MD). The lysate was incubated at 60 °C for 20 min to denature *E. coli* proteins and clarified by centrifugation at 30,000 x g for 30 min. The supernatant was then applied to a gravity-flow column containing a 0.5- to 1-ml bed volume of Ni-NTA agarose (Qiagen). The column was washed with 8 column volumes of wash buffer (20 mM Tris, 400 mM NaCl, and 10 mM imidazole, pH 8.0), and Csa3a was eluted in 10 mM Tris (pH 8.0), 50 mM NaCl and 300 mM imidazole. Csa3a was then applied to a calibrated Superdex 75 10/300 GL column (Cytiva) equilibrated with SEC Buffer (10 mM Tris and 50 mM NaCl pH 8.0) and fractions containing the major peak were combined. Protein concentrations were determined by Bradford assay [29] using Protein Assay Reagent (Bio-Rad) and bovine serum albumin as a standard. The purity and molecular weight of Csa3a were confirmed by SDS-PAGE.

2.3 Isothermal Titration Calorimetry

Protein samples were prepared by passage over a Superdex 75 (Cytiva) as described above. Lyophilized cA_4 was resuspended in the same SEC buffer used to prepare Csa3. These Csa3a and cA_4 stocks were centrifuged at 17,000 x g for 10 min to remove any aggregates, and the UV absorbance of the samples was measured to determine concentrations (ε_{280nm} for Csa3a = 8960 M^{-1} cm^{-1} and ε_{259nm} for cA_4 = 54,000 M^{-1} cm^{-1}), with samples prepared by dilution into the same buffer (25 μm Csa3a dimer, 250 μM or 500 μM cA4). In a MicroCal ITC200 (Malvern Panalytical), 250 μM or 500 μM cA_4 was injected (19 injections of 2 μL at 180 s intervals) into a cell chamber containing 207 μL of Csa3a dimer (25 μM). Isotherms were fit to a single-site model with MicroCal's ORIGIN software, assuming a single cA_4 binding site in each Csa3a dimer. Titrations were performed in duplicate. The averaged results are reported in Supplemental Table S1.

2.4 Cyclic Oligonucleotide Phosphodiesterase (Ring Nuclease) Assays

"Ring nuclease" assays were performed largely as previously described [30]. Briefly, unlabeled cA_4 was added to a final concentration of 150 μM to reactions containing 10 μM SsCsa3a in 10 mM Tris, 50 mM NaCl, pH 8.0, plus SUPERase•In inhibitor, and incubated at 50 °C. 10 μl aliquots were removed from the reaction mixture and quenched with equal volume phenol-chloroform at desired time points. Upon quenching, 20 μl H_2O was added to each aliquot to bring the sample volume up to 30 μl aqueous and 10 μl organic. 20 μl of the aqueous portion was then

removed and dried in a SpeedVac for 30 minutes. Samples were then re-suspended in 20 µl methanol, centrifuged at 16,000 x g for 10 min, and the top 15 µl was removed and again dried in a SpeedVac for 30 minutes. Finally, samples were re-suspended in 5 µl methanol, and spotted onto a glass-backed F254 silica gel TLC plate. The TLC plate was developed in 70% ethanol and 0.2 M ammonium bicarbonate (pH 9.3) until the mobile phase reached approximately 75% of the plate height. Plates were then dried and imaged with an F254 lamp and imaging box. The intensities of the residual cA_4 were quantitated using ImageJ image analysis software and normalized to time = 0. Using KDE LabPlot 2.8.2, the data were then fit to progress curves as described by Sternberg et al [31] to determine the first order rate constants (k_{obs}).

2.5 Crystallization and Data Collection

Purified Csa3a in SEC buffer was concentrated to 7.5 mg/ml with a 10,000 MWCO Amicon Ultra™ spin concentrator (Millipore). The protein was co-crystallized with cA_4 using hanging drop vapor diffusion. Drops were set up at 22 °C using 2 µl of Csa3, 2 µl of well solution (14–22% PEG MME 550, 100 mM imidazole pH 7.0, and 150 mM malate at pH 7.2), and 0.2 uL of 10 mM cA_4 (c-tetraAMP, BioLog Lifescience Institute). Co-crystals of Csa3a with cA_4 up to 0.30 mm × 0.075 mm × 0.075 mm in size were obtained in 3 to 5 days. Single crystals were flash frozen in liquid nitrogen and data was collected at 160 K using our MicroMax-007 X-ray generator and Rigaku R-axis 4^{++} image plate detector. Csa3-cA_4 crystals

diffracted to 2.45 Å resolution. Data were indexed, integrated and scaled in space group $P2_12_12_1$ (a = 46.1 Å, b = 93.4 Å, c = 105.2 Å) using the HKL2000 software package [32]. Data quality parameters are presented in Table 1.

Table 1 Data Collection

Data Set	SsCsa3-cA_4
Wavelength (Å)	1.54
Space Group	$P2_12_12_1$
Cell Constants (a,b,c; Å)	46.1, 93.4, 105.2
(α, β, γ; °)	90, 90, 90
Resolution Range[a] (Å)	27.91-2.45 (2.49-2.45)
Unique Reflections[a]	16,551
Average Redundancy[a]	10.4 (10.4)
I/σ[a]	36.4 (2.5)
Completeness (%)	96.8(100)
R_{sym}[a,b] (%)	7.9 (82.0)
R_{pim}[a] (%)	2.6 (26.5)
CC1/2[a]	0.99 (0.737)

[a]Numbers in parenthesis refer to the highest resolution shell.
[b]$R_{sym}=100*\Sigma_h\Sigma_i|I_i(h)-\langle I(h)\rangle|/\ \Sigma_h I(h)$ where $I_i(h)$ is the i^{th} measurement of reflection h and $\langle I(h)\rangle$ is the average value of the reflection intensity.

Table 1. Data Collection Parameters

2.6 Structure Determination and Refinement

The Phaser-MR module [33] of Phenix [34] was used to determine initial phases for Csa3a in complex with cA_4 by molecular replacement. The structure of apo-Csa3a (PDB ID: 2WTE) [6] was used as a search model. Anticipating conformational changes upon cA_4-binding, the N- and C-terminal domains were placed separately. 2 copies of each domain were placed, yielding a single dimer within the asymmetric unit (40.5% solvent). The molecular replacement solution

yielded a refined LLG of 2211, a TFZ score of 20.3, and an R-value of 42.4%. The structure was completed by iterative model building with Coot [35] and refinement with phenix.refine [34]. Temperature/libration/screw parameters [36] were included in the refinement, with each of the two Csa3a chains and the single cA_4 ligand divided into twelve temperature/libration/screw groups (chain A: 1–110, 111–157, 158–177, 178–112; chain B: 1–28, 29–55, 56–84, 85–97, 98–127, 128–157, 158–185, 186–212; cA4: 1–4). The final R_{work} and R_{free} were 20.3% and 24.30%, respectively. MolProbity was used for model validation [37], with 96% of residues falling in the most favored regions of the Ramachandran plot and none in disallowed regions. The overall MolProbity score [37] was 1.51, placing Csa3a-cA4 in the 99th percentile for overall geometric quality among protein crystal structures of similar resolution (2.458 ± 0.25 Å). Residue numbers in the model are consistent with the native, nontagged protein sequence. Residues 160–162, 174–176, 192–195, and 213–237 of chain A, and 213-237 of Chain B were not modeled due to lack of interpretable electron density. Additional details on model refinement and model quality are presented in Table 2. The model has been deposited in the Protein Data Bank (6W11). Figures were generated in Pymol [38]. To model DNA recognition in Figures 5 and 6, and Videos S1 and S2, subunit B was replaced with a superpositioned copy of subunit A to restore disordered residues of the wHTH domain.

Table 2 Model Refinement	
R_{work}[c] (%)	20.3 (28.8)
R_{free}[c] (%)	24.3 (34.6)
Real Space CC[d] (%)	84.8
Mean B Value (overall; Å^2)	85.0
Coordinate Error (based on maximum likelihood, Å)	0.37
RMSD from ideality:	
Bonds (Å)	0.005
Angles (°)	0.687
Ramachandran Plot[e]	
Most Favored (%)	96.0
Additional Allowed (%)	4.0
PDB Accession Code	6W11

[c]$R_{work} = \Sigma||F_o|-F_c||/\Sigma F_o|$ where F_o and F_c are the observed and calculated structure factor amplitudes used in refinement. R_{free} is calculated as R_{work}, but using the "test" set of structure factor amplitudes that were withheld from refinement (4.9%).
[d]Correlation coefficient (CC) is agreement between the model and $2mF_o$-DF_c electron density map.

[e]Calculated using MolProbity (35)

Table 2. Model Refinement Parameters

2.7 Electromobility Shift Assays (EMSAs)

EMSA fragments were generated by PCR using Phusion® DNA Polymerase, *S. solfataricus* P2 genomic template DNA, and the locus-specific primers listed in Supplemental Table S2. PCR products were the purified using the QIAquick PCR Purification Kit (Qiagen). 12 µL EMSA binding reactions containing 5.2 ng/µL DNA, varying concentrations of Csa3a, +/- 10 µM cA_4 were incubated at 37 ℃ for 30 min in TAE buffer. Samples were separated with a 2% w/v agarose gel, and bands

were visualized via ethidium bromide staining and imaged using an Alpha Innotech AlphaImager 2200 gel imager. Data were quantified using ImageJ image analysis software, where we followed the disappearance of the unbound DNA, and modeled with GraphPad Prism.

2.8 Fluorescence Polarization (FP) Assays

The 5'-FAM-labeled CAPPa (20 bp) and 'Idealized' CAPPa (28 bp perfect palindrome based on our CAPPa) were generated via incubation of complementary single stranded oligonucleotides (Supplemental Table S3) in 5 mM $MgCl_2$ at 94°C for 5 min followed by slow cooling to 21°C over ~3 hours. The 5'-FAM labeled 75 bp DNA fragment was generated via PCR using Phusion® DNA Polymerase (NEB), PH1-16 genomic DNA as template, and a 5'-FAM-tagged forward primer and untagged reverse primer (Supplemental Table S3). FP assays were conducted in 384-well black plates with optical bottoms (Nunc). Samples containing 30 nM 5'-FAM-labeled DNA probe and varied concentrations of Csa3a with or without 10 µM cA_4 were prepared in 20 mM SPG (2.5 mM succinate 8.75 mM NaH_2PO_4, 8.75 mM glycine), 50 mM NaCl, pH 6.5 with a total volume of 100 µL. Samples were covered to prevent evaporation and incubated at 21°C or 50 °C for 15 min. FP readings were recorded at 21 °C or 50 °C using a SpectraMax iD5 plate reader (Molecular Devices). mP values for the samples were calculated using a G factor of 0.314. Binding data were analyzed using GraphPad Prism.

2.9 Sulfolobus Strains, Growth Conditions

S. solfataricus PH1-16 (Δ*pyrEF*) was cultured in DT media (with 20 µg/mL uracil) at 78 °C and shaking at 65 rpm. An empty pSeSD1 vector and pSeSD1 containing *S. solfataricus* P2 csa3a (pSeSD-sso1445) were transformed into PH1-16 cells via electroporation [9, 39], and transformants were selected for via growth in DT media without uracil. Cultured transformants were plated on 0.7% GelRite DT media plates and incubated at 78°C until colony formation. A monoculture of transformed cells was started via inoculation of DT media with a single colony [9].

2.10 Acquisition Assay

Wild-type *S. solfataricus* PH1-16, PH1-16 containing an empty pSeSD vector, and PH1-16 containing the pSeSD-sso1445 vector were cultured in AT media (DT media with arabinose substituted for dextrose) until OD_{650} ~0.8. Overexpression of Csa3a was confirmed via protein purification from induced PH1-16 cells as for *E. coli* above. The first 2-5 leader proximal spacers of the CRISPR loci were PCR amplified using Phusion® DNA Polymerase (NEB), 3 µL of cell culture as the DNA template, and locus-specific primers (Supplemental Table S4). PCR products were separated on a 2% agarose gel and visualized via ethidium bromide staining.

3. Results

3.1 CRISPR/Cas Systems in S. solfataricus

We began our investigation with further analysis of CRISPR-Cas systems in *S. solfataricus*. The *S. solfataricus* genome harbors six CRISPR loci (A-F) containing

anywhere from 6-102 spacers, and at least seven cas gene clusters with varying levels of complexity. These include three Type I-A systems, three Type III-B systems, and a single Type III-D system (Supplemental Figure S1). Two Type I-A systems, denoted I-A(1) and I-A(2), are accompanied by genes for acquisition (*cas1*, *cas2*, *cas4* and *cas4a)* and pre-crRNA processing (*cas6*). However, with the exception of a potential phage-derived exonuclease/*cas4* present in the Type III-D cluster, the remaining clusters lack acquisition and maturation genes, and presumably rely on those associated with the I-A(1) and I-A(2) systems.

Within these clusters genes encoding CARF domain proteins are found associated only with the I-A(1) and I-A(2) systems. In the I-A(1) cluster these are csx1, crn1 (sso1393) and an unannotated CARF protein of unknown function (sso1397). The crn1 gene encodes a clade 7 CARF domain with CRISPR "ring nuclease" activity [4, 40]. CRISPR-ring nucleases degrade the cA_4 signal, providing an off switch for the cyclic oligoadenylate signal [40-43]. In the case of sso1393, this CARF domain ring nuclease is fused to a winged helix-turn-helix domain, reminiscent of the CARF-wHTH of Csa3. *S. solfataricus* contains a second crn1 gene, sso2081, 81 kb to the 5' end of the I-A(3) cluster. The Sso2081 Crn1 ring nuclease shows greater ring nuclease activity than Sso1393, and is thought to be the principal player in cA_4 degradation in *S. solfataricus* [40].

The I-A(2) locus harbors two additional CARF domain proteins, csa3b (sso1444) and csa3a (sso1445). *csa3a* is flanked on the 5' side by CRISPR locus C, which separates csa3b (sso1444) from csa3a (sso1445) and on the 3' side by a DDE-

family transposase, an acquisition cassette (*cas4*, *cas2*, *cas1*, and *cas4a*) and CRISPR locus *D*. Additionally, *csa3a* is in the reverse orientation relative to CRIPR locus C and the acquisition cassette. Overall, this suggests Csa3a may regulate expression of these elements [6].

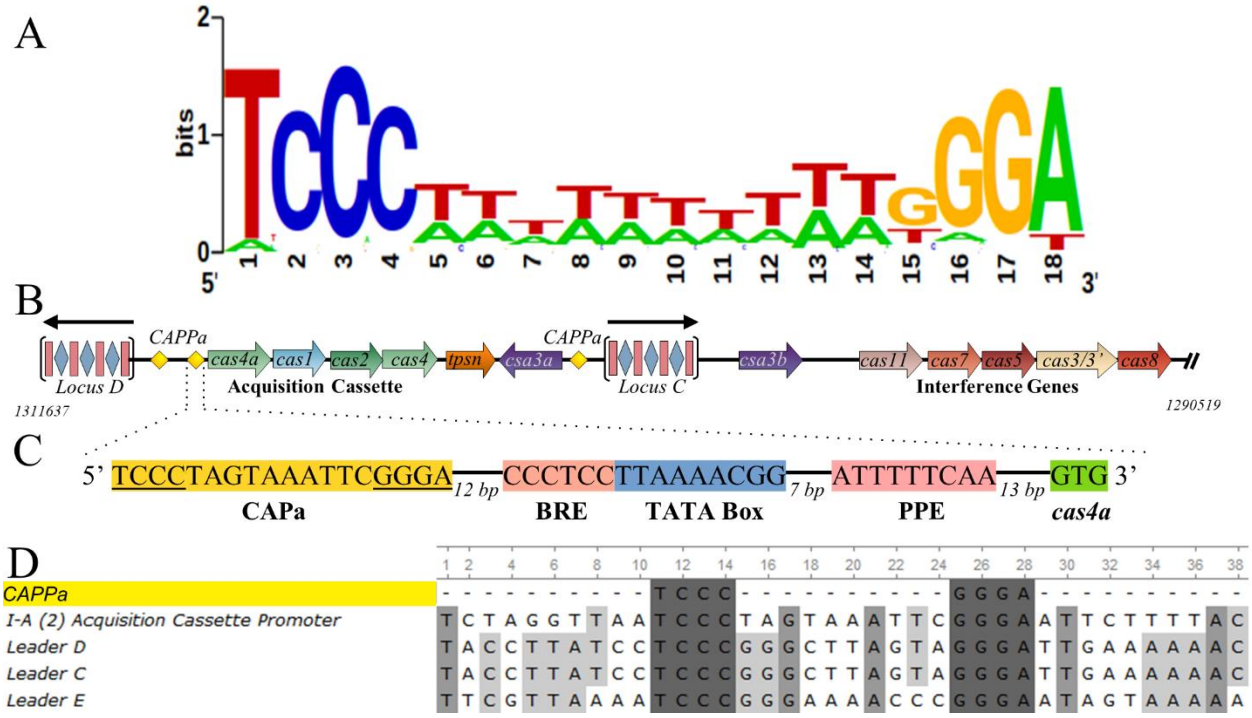


Figure 1. CRISPR Associated Palindromes in S. solfataricus. (**A**). Sequence logo for the CRISPR-associated palindrome CAPPa ($TCCCN_8GGGA$). **B**) Organization of 1-A(2) acquisition cassette and adjacent CRISPR loci. CAPPa motifs (yellow diamonds) are found at four locations in the *S. solfataricus* genome; upstream of the 1-A(2) acquisition cassette as well as CRISPR loci C, D and E. **C)** CAPPa is immediately upstream of the B recognition element (BRE), the TATA box and the promoter proximal element (PPE). **D)** CAPPa sequences in *S. solfataricus* P2.

3.2 CRISPR Associated Palindromes

Relative to *S. solfataricus*, *S. islandicus* Rey15A harbors a smaller ensemble of CRISPR loci and CRISPR-associated (cas) genes. Liu *et. al.* localized a Csa3a

binding site in the promoter of the Rey15A acquisition cassette, and expanded on that in 2017 by reporting 24 bp pseudo-palindromic binding sites in the leader sequences of the two CRISPR loci as well [23, 25]. They also identified similar sequences in *S. solfataricus* upstream of the I-A(2) acquisition cassette and in the leader sequences of the C, D and E CRISPR loci. Similarly, our own analysis identified these same sequences. Importantly, when we analyzed these sequences with MEME, we identified a consensus 18-bp palindromic motif that is conserved in 27 members of the order Sulfolobales (Figure 1A). Further, this motif is generally absent from the promoters of non-CRISPR associated genes in *S. solfataricus*, i.e., it is a CRISPR-Associated Promoter Palindrome (CAPP). To differentiate this motif from an unrelated binding site reported for Csa3b [24], we refer to this motif as CAPPa. The CAPPa element is present 12 bp upstream of the BRE and TATA box for the I-A(2) associated acquisition genes [25], consistent with a role in transcriptional regulation of the acquisition cassette, as well as the leader sequences of CRISPR loci C, D and E (Figure 1). Last, we also find a nearly exact correspondence between the occurrence of Csa3a and the CAPPa motif in these genomes (Supplemental Table S5 and S6).

3.3 Csa3a Binds cA_4

We next sought to confirm the hypothetical interaction between Csa3a and cA_4. Indeed, isothermal titration microcalorimetry (ITC) at 25 °C and 250-500 µM cA_4 gave a dissociation constant (K_d) *of* 1.0 to 1.5 µM depending on the cA_4

concentration (Figure 2, Supplemental Table S1), Notably, while the expected stoichiometry between the Csa3a dimer and cA_4 is one to one, the ITC data shows a lower than expected stoichiometry of approximately 0.6 across all titrations. However, sample centrifugation (17,000 x g, 10 min) after ITC resulted in a distinct protein pellet, and control experiments with matching buffer in the absence of cA_4 also showed similar pellet sizes, indicating the presence of cA_4 does not affect the aggregation observed in the experiment. In contrast, Csa3 left at 25°C without stirring did not give a pellet. Thus, the lower stoichiometry is likely due to protein aggregation from the agitation innate to the ITC assay. Importantly, ITC analysis also demonstrated that, at 25°C, the binding event is endothermic, and is thus entropically driven.

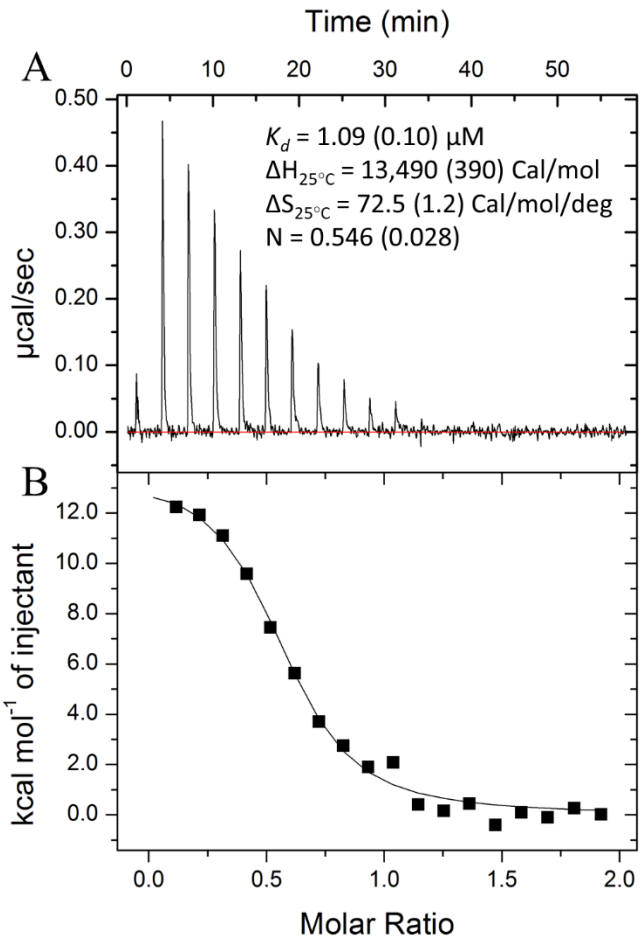


Figure 2. ITC analysis of the Csa3-cA4 interaction. **A**) Differential power recorded over the course of the ITC experiment for titration of 250 µM cA_4 into 25 µM Csa3a dimer. **B**) Integration of areas under peaks corresponding to energy absorbed upon addition of cA_4 to Csa3a as a function of molar ratio [cA_4]/[Csa3a dimer].

3.4 Csa3a Lacks Ring Nuclease Activity

CARF domain proteins, like Sso1393, frequently exhibit low level ring nuclease activity that attenuates their activation by cA_4 [40, 43-45]. To determine if Csa3a possesses ring nuclease activity and is potentially self-regulating, unlabeled cA_4 was incubated with Csa3a for up to 33 hours (Figure 3). We observed no cA_4 degradation, suggesting the Csa3a CARF domain acts as a signal receptor, but is

not involved in clearing the cA_4 signal. In this regard, we note that Csa3a lacks a putative catalytic lysine residue found in many CARF ring nucleases [40].

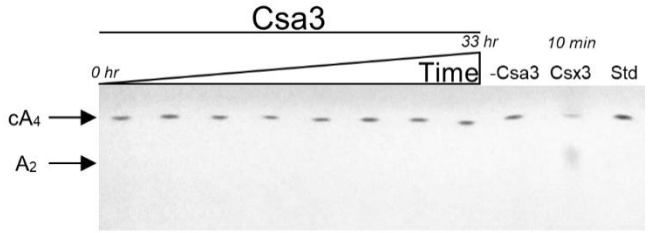


Figure 3. Csa3a does not exhibit ring nuclease activity. **A**) TLC analysis showing a representative time course for the degradation of unlabeled 150 μM cA_4 incubated with 10 μM Csa3a dimer at 50°C. Lanes 1-8 are time points at 0, 1, 2, 4, 6, 8, 16, 24, and 33 hours. Lane 9 was a 33 hour incubation in the absence of Csa3a. Lane 10 was a 10 minute incubation with 10 μM AfCsx3 dimer and 200 μM Mn^{2+} to serve as a positive control [30]. Lane 11 was loaded with 1.5 nmol cA_4 standard.

3.5 Csa3a Binds cA_4 in the CARF Binding Pocket

To elucidate the molecular basis for cA_4 recognition by Csa3a and any conformational changes induced by cA_4 binding, we co-crystallized Csa3a with cA_4 and determined the structure by X-ray crystallography. While the crystallization conditions were significantly different from those reported for the cA_4-free structure of Csa3a first reported by Lintner et al [6], it crystallized in the same space group with similar unit cell parameters (Table 1). Consistent with this crystal form, the asymmetric unit contains a single Csa3a homodimer, and the overall structure is largely similar to the cA_4 free structure. However, the position of the wHTH domains relative to the CARF domains are noticeably different, and these larger conformational changes appear to be driven by subtle conformational changes in the N-terminal CARF domain associated with cA_4 binding.

The N-terminal CARF domain retains its overall fold, a doubly wound, mixed ß-sheet with flanking α-helices. Similar to a classic dinucleotide binding domain, the first five ß-strands ($ß_1$-$ß_5$) run parallel and are connected by helical, righthanded crossovers. However, the fold then diverges from the classic "Rossmann fold", as $ß_5$ makes a reverse turn into $ß_6$, which runs antiparallel to $ß_5$, giving the sheet a $ß_3$(↑)-$ß_2$(↑)-$ß_1$(↑)-$ß_4$(↑)-$ß_5$(↑)-$ß_6$(↓) topology. The reverse turn into $ß_6$ also results in the loss of a covering alpha helix. Critically, CARF domain dimerization restores this structural element, placing $α_4$ of the second subunit across the $ß_5$/$ß_6$ end of the ß-sheet. The $α_4$ helices are thus buried at the dimer interface, where they run parallel to each other. Overall, the non-canonical $ß_5$/$ß_6$ reverse turn and loss of the connecting α-helix is critical to dimer formation, and are thus defining features of the CARF domain fold [6, 30].

As expected [6], a single cA_4 molecule is bound within a 2-fold symmetric pocket spanning the subunit interface at the "base" of the Csa3a dimer (Figure 4, Supplemental Video, Supplemental Figure S2). The floor (or ceiling, depending on orientation) of this pocket is formed by neighboring $ß_4α_4$ loops from each of the two subunits. Principal residues in this loop are Gly96 and Arg98. The Arg98 side chains run antiparallel to each other as they reach across to the subunit interface and hydrogen bond to the main chain carbonyls of Gly96 in the adjacent subunit. The unsaturated Gly96 main chain amines and the Arg98 side chains then mediate alternating interactions with the phosphate groups of cA_4. These interactions explain the strong conservation of these residues in "motif 2" of Lintner et al [6].

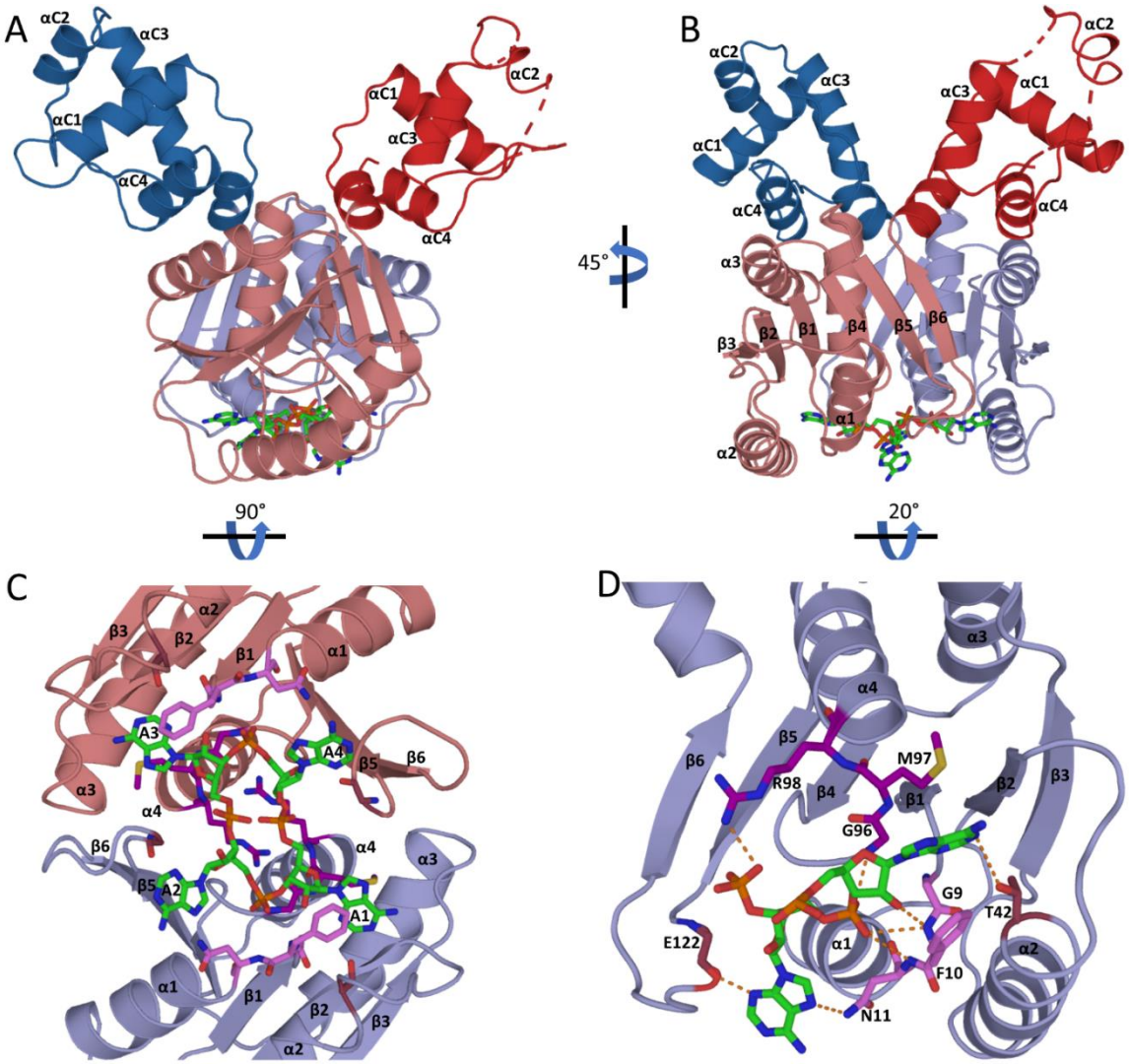


Figure 4. The Csa3a CARF domain binds cA4. A**)** Ribbon diagram of the Csa3a dimer bound to cA_4. The N-terminal CARF domains that form the cA_4 binding pocket are colored light red/blue, the C-terminal wHTH domains are colored dark red/blue, and cA_4 is shown in a stick representation with C atoms colored green. A single cA4 molecule is bound in the 2-fold symmetric ligand binding site, across the dimer interface of the CARF domain. **B.** The cA_4 -bound Csa3a dimer rotated 45° about the vertical axis shows the location of cA_4 in the binding pocket relative to secondary structural elements in chain B (red). **C.** Ribbon diagram of the cA_4 -bound Csa3a dimer looking into the cA_4 binding pocket. Relative to A, the structure is rotated 90° about the horizontal axis. Motif 1 (pink), Motif 2 (purple), T42 (deep red), and E122 (deep red) which make direct interactions with cA_4 are represented as sticks. **D.** Ribbon diagram of a single Csa3a subunit bound to bases 1-2 of cA_4. The structure is rotated -20° about the horizontal axis relative to B. Hydrogen bonds are shown as orange dashes.

cA_4 is bound in an oval-shaped conformation with all bases present in the anti-conformation, radiating outward from the oval-shaped ribose-phosphate backbone. Further, each ribose is present in a 2'-endo confirmation. Consistent with this, Molprobity [37] identifies the dihedral backbone conformations for nucleotides 2 and 4 at the distal ends of the oval as 6p conformers, and assigns nucleotides 1 and 3 as "outliers" of the 2o conformer [46]. Details for the RNA conformers are found in Supplemental Table S7. The $\beta_1\alpha_1$ and $\beta_5\beta_6$ loops project above the floor of the cA_4 binding site in an alternating pattern about the dimer axis, where they successively interdigitate between each of the four adenosine moieties. Specifically, the $\beta_1\alpha_1$ loop inserts between bases 1 and 2 in subunit A, and between bases 3 and 4 in subunit B. Similarly, the $\beta_5\beta_6$ loop in subunit A rises between bases 2 and 3, and the same loop in the B subunit between bases 1 and 4 (Figure 4).

The first nucleoside is thus bound in a groove with walls formed by the $\beta_5\beta_6$ loop of subunit B and the $\beta_1\alpha_1$ loop of subunit A. The conserved Gly9-Phe10 in the $\beta_1\alpha_1$ loop [motif 1 of Lintner et al, [6]] play a pivotal role. The main chain NH of Phe10 hydrogen bonds to O2' of the ribose and the edge of the adenine base packs against the main chain of Gly9, allowing the Phe10 side chain to extend over the adenine ring. Val39 of the $\beta2\alpha2$ loop also packs on top of the adenine ring, while the Met97 side chain ($\beta_4\alpha_4$ floor loop) provides a foundation underneath the adenine to complete a hydrophobic binding pocket. Except for Thr42 in the $\beta_2\alpha_2$ loop, which H-bonds to N1 of the adenine moiety, base specific interactions are not obvious.

In contrast to the first nucleoside, the second adenosine is bound entirely within subunit A, between the $\beta_1\alpha_1$ and $\beta_5\beta_6$ loops. The adenine ring settles on top of main chain atoms of the $\beta_5\beta_6$ loop (residues 122-124) and Phe14 of the $\beta_1\alpha_1$ loop, leaving the opposite face of the adenine solvent exposed. In addition, Thr13 accepts an H-bond from the adenine -NH_2 while the Asn11 side chain donates a hydrogen bond to N7 of the adenine ring. Finally, the Asn11 main chain NH hydrogen bonds to the 5'phosphate of this nucleotide.

Overall, then, the first subunit plays a predominant role in recognition of the first two nucleotides of cA_4. Similarly, the second subunit largely mediates recognition of the last two nucleotides, using a similar set of interactions. However, the cA_4 is not bound in a completely symmetric fashion. Noticeably, a crystal contact with a neighboring asymmetric unit results in a cation-π interaction between Arg200 and the solvent exposed adenine ring of the fourth nucleotide. This interaction appears to pull the adenine out of its binding pocket, breaking the interactions with Asn11 and Thr13. Further, the $\beta_5\beta_6$ loop near the crystal contact is also found in an alternate conformation relative to subunit A. Finally, each of the ribose moieties are present in a 2'-endo conformation and for residues 1, 2 and 3, the 2'-OH is hydrogen bonded to either O1P or O2P of the neighboring phosphate group. While the ribose in the 4^{th} nucleotide is also 2'-endo, it reaches across the cA_4 molecule to instead hydrogen bond to the O2P of the 3^{rd} nucleotide. This is possible, in part, because the phosphate group linking the 2^{nd} and 3^{rd} nucleotides does not sit down on the floor of the pocket, as it does for the 5'-phosphate of the first nucleotide.

Whether these differences are the result of crystallographic contacts, or there is inherent asymmetry in the solution bound form is not clear.

3.6 cA_4 Induced Conformational Changes

Local to the cA_4 binding site, the most prominent conformational changes are modest movements of the $\beta_1\alpha_1$ and $\beta_2\alpha_2$ loops as they move away from the dimer axis, expanding the conserved pocket to accommodate cA_4. At the same time, there are significant side chain rearrangements as Phe10 rotates over the adenine ring of the first (and third) nucleotide. These movements are accompanied by a small, yet significant rotation of the $\beta_1\alpha_1\beta_2\alpha_2\beta_3\alpha_3$ end of the CARF domain (Figure 5, Supplemental Video S1). Critically, the α_3 helix runs antiparallel to the C-terminal helix of the winged helix-turn-helix DNA binding domain. This small rotation of the α_3 helix is, in turn, transmitted to the C-terminal helix of the wHTH domain, acting as a pivot point for an approximate 7° rotation of the wHTH domain relative to the CARF domain. Relative to the Csa3a dimer, the net result is a twisting movement of the wHTH domains towards each other, into a conformation that would appear more favorable for binding linear B-form DNA. However, superpositional DNA docking using the same dimeric OhrR/DNA complex (PDB code: 1Z9C) originally used by Lintner et al. [6] suggests additional movement of the wHTH domains or some DNA bending might still be required. (Figure 5, Supplemental Video S2).

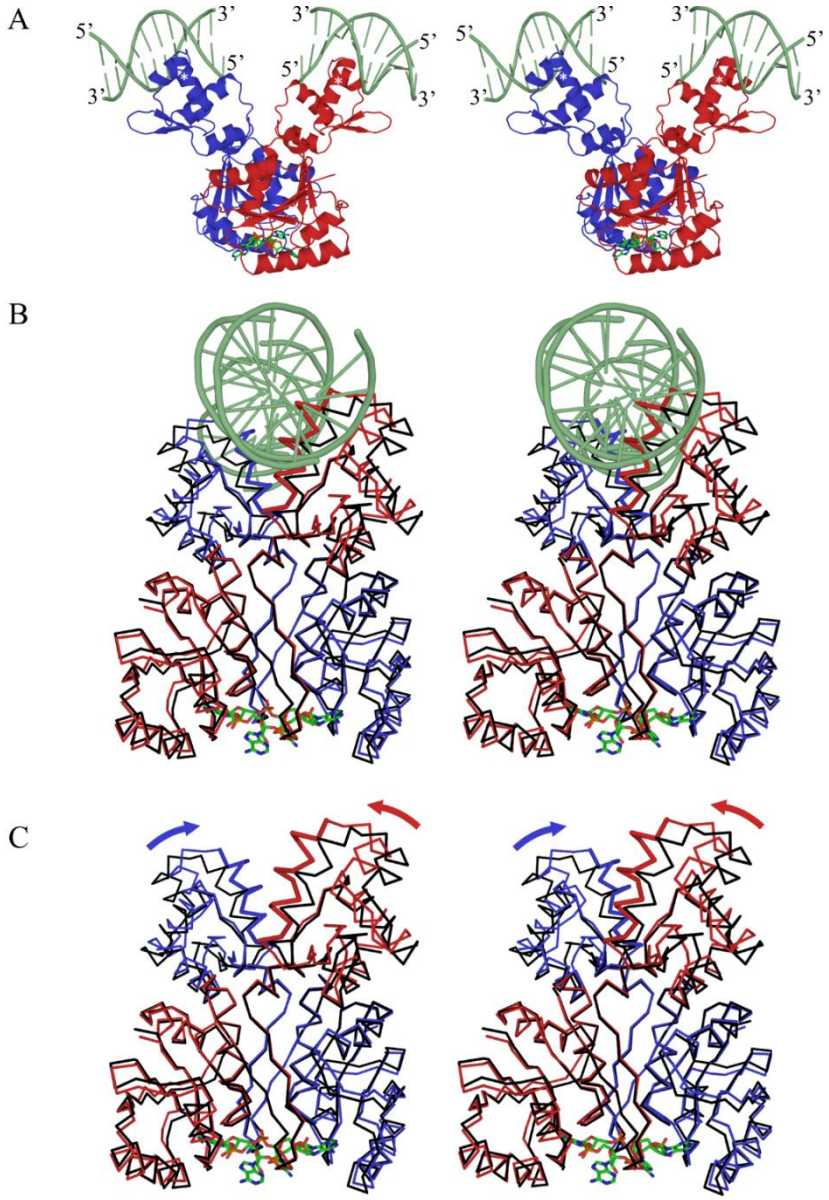


Figure 5. cA4 induced conformational changes. **A.** Divergent stereo image of the Csa3a/cA4 complex docked to DNA (light green) by superposition of individual wHTH domains onto *B. subtilis* OhrR (PDB code: 1Z9C [47]) (RMSD: 1.7 Å). We were unable to superposition the Csa3a dimer onto a single strand of B-form DNA, suggesting further conformational changes in Csa3a or DNA bending might occur. **B.** Significant movement of the wHTH domains is seen at the putative protein/DNA interface in the Csa3a/cA_4 complex (blue and red) relative to the unbound structure (black). **C.** The rotation of the wHTH domains (direction emphasized by arrows) relative to the CARF domain is a result of modest conformational changes in the CARF domain.

Our previous multiple sequence alignment did not differentiate between Csa3a and Csa3b [6]. While this alignment nicely identified the conserved cA_4 binding site, considering the different DNA sequences recognized by *S. islandicus* Csa3a and Csa3b [24, 25, 48], it might not be expected to identify conserved residues involved in DNA recognition. We thus constructed a new multiple sequence alignment with 21 Csa3a sequences. The alignment identifies three strictly conserved residues in the recognition helix – E173, K174, and N178 – that are positioned to make base-specific interactions. Additional conserved resides positioned to interact with the ribose-phosphate backbone include E146, E147, K171, S172, T175, K179, E182, K184, G193, D195, and R196 (Figure 6).

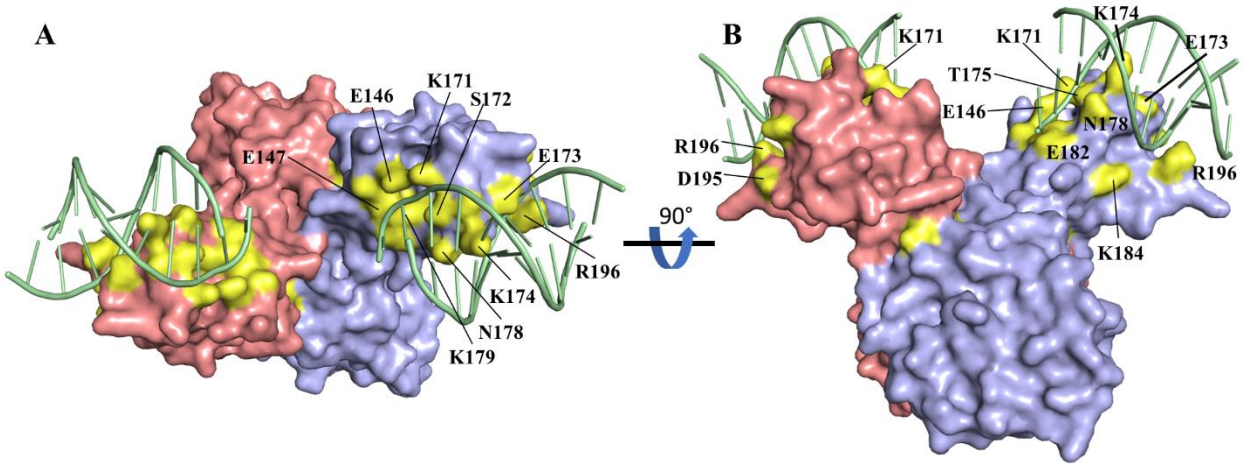


Figure 6. Conserved residues in Csa3a wHTH domain. A) Surface model of Csa3a with strictly conserved residue in the wHTH domain in yellow and DNA in orange. B**)** Rotated 90° about the horizontal axis. The majority of the conserved surface residues in the wHTH domain are located at the putative protein-DNA interface.

3.7 Csa3a Binds CAPPa Nonspecifically, in a Cooperative and cA_4 Independent Manner

We first attempted to demonstrate cA_4-modulated recognition of the putative Csa4 promoter for the *S. solfataricus* I-A(2) acquisition cassette using an electromobility shift assay (EMSA). We found that Csa3a will indeed shift a 350 bp fragment incorporating the putative promoter with relatively high affinity, giving a measured K_d of approximately 250 - 300 nM (Figure 7). Interestingly, the binding curve was sigmoidal, suggesting DNA binding was cooperative, and indeed, the fit to the observed data improved when the Hill coefficient was not constrained. Similar results were found for fragments containing the conserved CAPPa motifs in the leader sequences of CRISPR loci C, D and E as well (Supplemental Figures 3-5). When cA_4 was included in the assay, however, affinity for the putative promoter region was not increased. In addition, "negative" controls, that lacked the palindromic motif, BRE and TATA binding sites bound with similar affinity, indicating the observed DNA-binding is nonspecific. Consistent with non-specific binding and the length of the target DNA, we observed a spectrum of shifted products, indicative of multiple binding events per DNA strand.

We considered whether the affinity of Csa3 for cA_4 ($K_d \sim 1$ μM) might be insufficient for the EMSA assay. Specifically, that if cA_4 dissociates from Csa3, it is pulled in front of the Csa3/DNA complex during electrophoresis, separating it from the complex and changing the affinity of Csa3 for DNA. We thus developed a fluorescence polarization (FP) assay in which the cA_4 concentration remains

constant. This assay also allowed us to work at temperatures as high as 50 °C, and at biologically relevant pH (pH = 6.5, [49, 50]). Preliminary work at 21°C and 50 °C indicated binding affinity and specificity were not affected by temperature. We subsequently chose to work at 50 °C, as this temperature is closer to the optimal growth temperature for *S. solfataricus* (80 °C). Additionally, changing pH from 8.0 to 6.5 did not affect binding affinity or specificity.

We first examined the affinity of Csa3a for a 20 bp fragment centered on CAPPa in the I-A(2) cas4 promoter. Surprisingly, the affinity for this smaller 20 bp fragment was only 3.2 +/- 0.6 μM, much less than for the 350 bp fragment used in the EMSA assay (Figure 8). And critically, a statistically significant increase in affinity in the presence of cA_4 (K_d = 2.7 +/- 0.6 μM) was not observed. We also observed binding was no longer cooperative, which might be explained by the smaller DNA fragment (20 bp) which is able to accommodate only a single Csa3 dimer. Indeed, when we instead used a 75 bp fragment, extending in the upstream direction away from the BRE and TATA elements, cooperative binding was restored, along with the higher affinity (K_d = 270 +/- 20 nM) previously observed in the EMSA assays. However, cA_4 again failed to provide a significant increase in affinity (K_d = 210 +/- 20 nM). We thus concluded that *S. solfataricus* Csa3a binds CAPPa nonspecifically, in a cA_4 independent manner. This is clearly at odds with a simple model wherein cA_4 increases the affinity of Csa3a for CAPPa, Csa3a then binds CAPPa, with subsequent recruitment of TFB, TBP and RNA polymerase. Given the data supporting transcriptional control of the acquisition cassette and

CRISPR loci in *S. islandicus* [23, 25], alternative models for the initial high affinity interaction involving additional proteins might need to be considered.

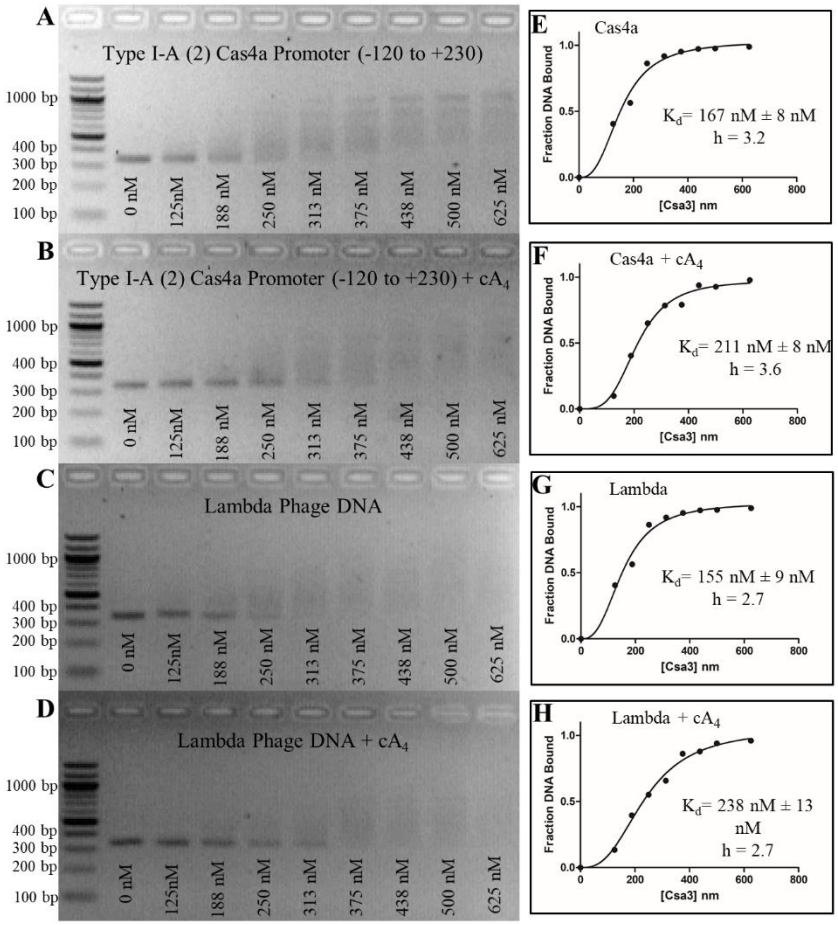


Figure 7. Nonspecific binding to the Type I-A(2) cas4a promoter. A-D) Csa3a binds both the cas4a promoter region containing CAPPa and the control Lambda DNA fragments with 150 - 250 nM affinity, indicating a nonspecific interaction in the absence of cA_4. The addition of cA_4 does not affect binding affinity or specificity. EMSA experiments were performed with 18 nM DNA and increasing [Csa3] dimer (denoted below the lanes). E-H**)** Graphs generated in GraphPad Prism display curves for specific binding with a Hill coefficient (h) fit to binding data generated using ImageJ image analysis software. EMSA fragments are detailed in Supplemental Table S4.

In this light, the cooperative nature of the non-specific binding remains of interest. Our modeling suggested the 20 bp fragment was sufficient to allow binding by the Csa3 dimer. But to ensure the target CAPPa sequence was indeed long enough to accommodate the full span of interactions, we repeated the experiment with a 28bp DNA fragment centered on an idealized CAPPa motif (Supplemental Table S3). This intermediate length fragment showed affinity (K_d = 1.3 +/-μM) close to that of the 20 bp CAPPa, again with non-cooperative binding. We did, however, observe additional low affinity binding at higher ligand concentrations, indicative of a second, even lower affinity binding site. This mechanism was not observed for either the 20 bp or 75 bp fragments. We rationalize this as binding of a single Csa3a subunit to DNA extending from either side of the "higher affinity" Csa3/DNA complex. In this light, we conclude the nearly 10-fold difference in affinity between the 20 and 75 bp targets is a result of the cooperative binding of two dimers, leading to increased affinity compared to the 20 and 28 bp fragments, which allow binding of only a single dimer.

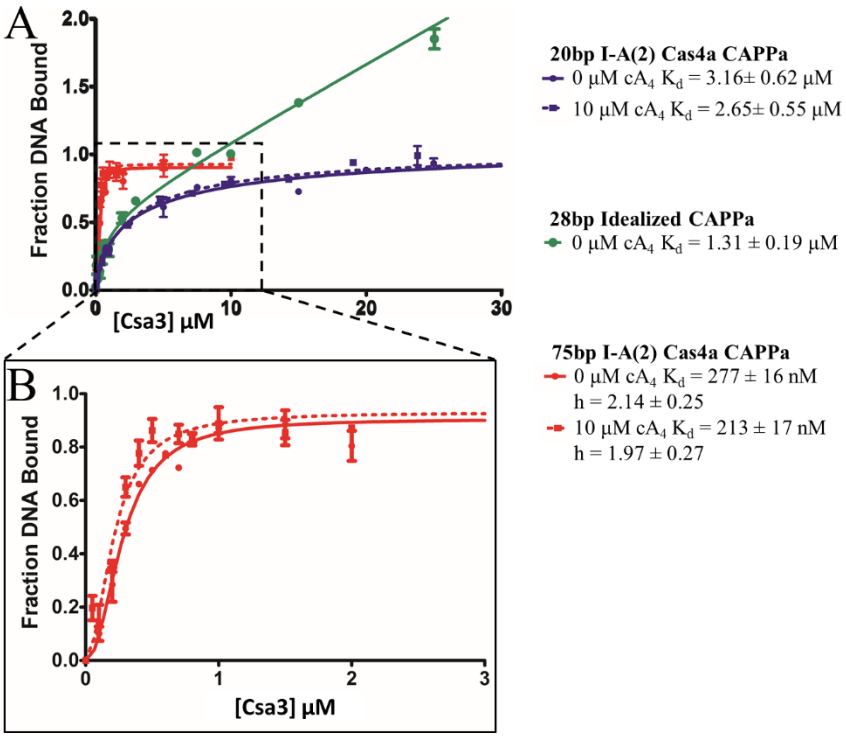


Figure 8. Csa3a binds CAPPa nonspecifically, in a cooperative and cA4 independent manner. A) DNA binding was measured by fluorescence polarization at 50 °C. Both the 20bp I-A(2) cas4a CAPPa (blue) and 28 bp idealized CAPPa (green) bound with low micromolar affinity, while the 75 bp I-A(2) cas4a CAPPa (red) was bound with 200-300 nM affinity. The addition of 10 µM cA_4 (dotted lines) did not significantly change the affinity. Linear nonspecific binding after saturation was observed for the 28 bp idealized CAPPa. **B**. Cooperative binding with a Hill coefficient of ~2 was observed for the 75 bp I-A(2) Cas4a CAPPa fragment. ssDNA oligonucleotides/primers used for generation of FP probes are listed in Supplemental Table S4.

3.8 Csa3a Overexpression Does Not Stimulate Spacer Acquisition in S. solfataricus

Liu et al. demonstrated that Csa3a overexpression stimulates spacer acquisition in *S. islandicus* [23]. We duplicated this experiment in *S. solfataricus* by transforming PH1-16 cells with a pSeSD1 vector containing SsCsa3a (pSeSD-sso1445), inducing overexpression with arabinose, and PCR amplifying the leader

proximal regions of the CRISPR loci. The PH1-16 cells are an *S. solfatricus* P1 ΔpyrEF variant with CRISPR-Cas gene content and organization that is highly similar to *S. solfataricus* P2. The Csa3 proteins in these strains show 100% identity at the amino acid level. Interestingly, we did not observe new spacer acquisition upon Csa3 overexpression in this second *Sulfolobales* (Figure 9). To confirm the negative result was not due to a heterogeneous culture of transformed and untransformed cells, we repeated the experiment with a monoculture started from a colony of transformed PH1-16 cells grown on DT GelRite plates with the same result. In addition, we also affinity purified the His-tagged Csa3a from these same cells, confirming that Csa3a was indeed expressed.

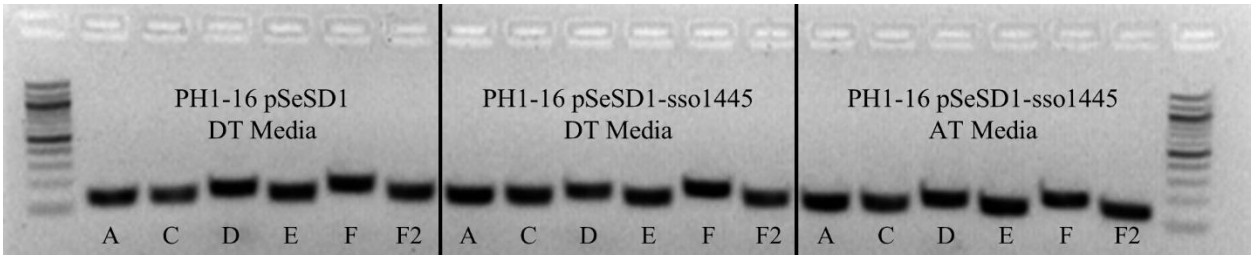


Figure 9. Overexpression of Csa3a in S. solfataricus does not lead to spacer acquisition. Leader proximal regions of the CRISPR loci were amplified from *S. solfataricus* PH1-16 cells carrying the empty pSeSD1 plasmid grown in DT media (uninduced, left), cells carrying the pSeSD1-sso1445 Csa3a expression plasmid grown in either DT media (uninduced, middle) or AT media (arabinose induced, right). Acquisition of a new spacer would appear as a secondary band ~60 bp larger than the band amplified from WT cells. CRISPR loci are labeled below the bands. F2 is the second half of locus F which is separated from the leader proximal region (Supplemental Figure S1).

4. Discussion

4.1 Physiological Implications of Low Micromolar cA_4 Affinity

While cyclic oligoadenylate signals have been identified for a number of CARF domain proteins, including those with RNase (Csx1/Csm6) and DNase effector domains (Can1/Card1 [51]), as well as single domain ring nucleases (Crn1, Crn3 [18, 30, 40]), to our knowledge this is the first formal demonstration that cA_4 is recognized by CARF domain transcription factors with wHTH effector domains. We were initially surprised to find the affinity of Csa3a for cA_4 is significantly lower than that reported for these other CARF domain proteins [30, 51], leading us to ask if cA_4 recognition is physiologically relevant. In this regard, Athukoralage et al. have modeled cellular cA_4 levels over the course of viral infection [42]. Their work suggests that detection of a single viral RNA by Type-III systems results in a burst of cA_4 production, potentially leading to cellular concentrations on the order of 6 μM cA_4. In this light, 1.1 micromolar affinity for cA_4 could indeed be physiologically relevant. It is also consistent with evidence that CRISPR systems have evolved to avoid deleterious levels of autoimmunity by limiting the rate of spacer acquisition [23], and that increased rates of spacer acquisition in mutants have been shown to lead to higher levels of toxicity [2]. The relatively low affinity of Csa3a for cA_4 may reflect fine tuning to balance the benefits of protection against a deleterious autoimmune response [23]. In this scenario, acquisition is a tightly regulated event, that is activated only when type III systems have detected invading mRNA, and

existing type I systems have failed, leading to sustained production of cA_4 by the type III systems.

4.2 cA_4 Induced Conformational Changes

*T*he structure of the Csa3a-cA_4 complex is among the first CARF domain structures to show significant conformational changes upon cA_4 binding. This may reflect the lack of ring nuclease activity. As expected, the conserved motifs of Lintner et al. [6] play significant roles in anchoring cA_4 in the binding pocket. Local conformational changes in the CARF domain binding pocket upon cA4 recognition then propagate globally to move the wHTH domains into positions that appear more favorable for DNA recognition. Specifically, cA_4 binding pushes on the sides of the binding pocket, causing a relative rotation of a core α_1, β_2, α_2, β_3, α_3 unit within each CARF domain. This movement, in turn, results in a pivot of the wHTH domains about the α_3 helix of the CARF domains, as the wHTH domains move toward each other (Supplemental Videos S1 and S2). Though movements of the Csa3 effector domains are more modest, these movements are reminiscent of those seen in CARD1 [51], and collectively these structures may define features common to the activation of CARF domain proteins by cA_4.

4.3 cA_4 Recognition is Entropy Driven

The endothermic binding observed by ITC indicates an entropy driven process, at least at 25 ^{o}C. Notably, the adenine bases in the cyclic oligoadenylates are solvent exposed and unable to base stack. Clathrate-like water structures are

thus expected on each side of the solvent exposed bases in the unbound ligand. Upon binding, the bases are desolvated and the ordered water is released, which is an entropically favorable process. In addition, the cyclic nature of the ligand suggests it is largely pre-ordered for binding, as is the binding pocket. Thus, there is a relatively small entropic penalty for immobilization of the ligand in the binding pocket. Overall, the increased disorder in the released solvent is greater than the increased order in the protein and ligand, resulting in an entropy driven process for recognition of cA_4 by Csa3. Importantly, these observations suggest entropy will play a significant role in driving cOA binding at thermophilic temperatures, and for the higher affinity interactions of other CARF domain proteins as well.

4.4 *Transcriptional Regulation of Spacer Acquisition and CRISPR Loci by Csa3*

Evidence that Csa3a regulates expression of the acquisition genes and plays a role in regulating the uptake and expression of new spacers in *S. islandicus* is quite strong. In particular, overexpression of Csa3a in *S. islandicus* Rey 15A increases transcription of the acquisition cassette and the incorporation of new spacers, and Csa3a has been observed to bind the *S. islandicus* cas4 and cas1 promoters in vivo [22, 26]. It would be surprising then, if Csa3a were not involved in regulating spacer uptake in other members of the Sulfolobales as well, especially since there is strong synteny among type I-A(2)-like acquisition cassettes in the Sulfolobales (Supplemental Table S5), and the presence or absence of Csa3a in these systems correlates nearly exactly with the presence or absence of CAPPa in the Cas4a promoters (Supplemental Tables S5 and S6). At the same time, however, Csa3a

overexpression in *S. solfataricus* fails to stimulate spacer acquisition, indicating distinct difference between *S. islandicus* and *S. solfataricus*. These differences are perhaps not so surprising as CRISPR-Cas in *S. solfataricus* is significantly more complex with regard to the number of cas genes and CRISPR loci. Further, in contrast to activation by Csa3a, *S. islandicus* Csa3b instead represses spacer acquisition [52]. How overexpression of Csa3a overcomes repression by endogenous Csa3b in *S. islandicus* is unknown, as is the state of cA_4 production under these experimental conditions. The lack of spacer acquisition in *S. solfataricus* upon overexpression of Csa3a may simply indicate Csa3a is unable to counter repression by Csa3b or other regulatory elements. Clearly, critical details remain to be elucidated in both *S. islandicus* and *S. solfataricus*.

Overall, we believe transcriptional regulation by Csa3a in response to the cA_4 second messenger remains an attractive hypothesis, one that is strengthened by observation of the cA_4/Csa3a interaction and the associated conformational changes in the structure of Csa3a presented here. At the same time, our data rule out a simple model wherein cA_4 produced by Cas10 in one of the type III surveillance complexes activates Csa3a binding to the CAPPa motif, followed by recruitment of transcription factor B (TFB), TATA binding protein (TBP) and RNA polymerase. A more complicated model might also be indicated by the non-canonical TATA box in the acquisition cassette promoters reported by Liu et al [25]. Nonetheless, a specific high affinity interaction with the promoter elements is likely to require the presence of additional factors, and to be a cooperative event.

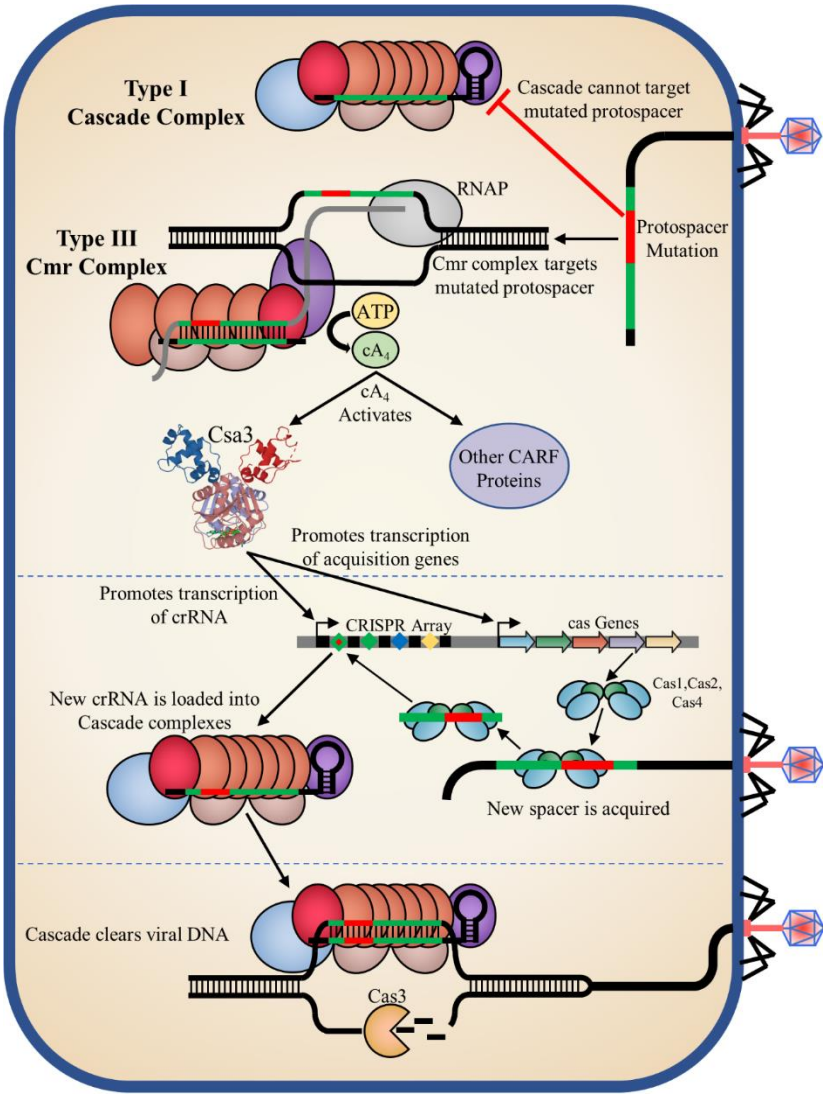


Figure 10. Model for an integrated Class 1 immune response. Top) Infection with a virus harboring an inexact match to archived CRISPR spacers may hinder recognition by PAM-dependent Type I systems, allowing the infection to progress to transcription of viral mRNA. The more promiscuous PAM-independent Type III interference complexes may still recognize the mRNA target and produce cA_4. cA_4 activates Csa3a (and potentially Csa3b), which promotes transcription of acquisition genes, resulting in spacer acquisition. Csa3a also promotes transcription of new crRNA's which are loaded into aCascade surveillance complexes, restoring recognition of the invading viral DNA to the Type I systems.

4.4 An Integrated, Synergistic Class 1 CRISPR-Cas Immune Response

Regardless of the specific mechanisms of transcriptional activation and repression by Csa3a and other Csa3 family members, confirmation of cA_4 as a Csa3a ligand in *S. solfataricus* now provides a critical, direct link between the Type-III surveillance complexes that synthesize cA_4 upon detection of invading nucleic acid [53], and Csa3 family transcription factors that regulate acquisition gene expression, transcription of CRISPR loci, DNA repair, and potentially the transcription and regulation of Type-I-A surveillance complex activity [6, 23-25, 54]. Thus, the general pathway for Csa3a-mediated spacer acquisition can still be inferred (Figure 10). For example, consider a cell infected with virus containing a protospacer mutation in the PAM sequence, rendering it un-targetable by the Type-I-A Cascade systems. The mutated protospacer will still be targeted by the more promiscuous and PAM independent Type-III surveillance complexes. Upon recognition of the mutated protospacer, the Cas10 subunit in the Type-III surveillance complex will synthesize cA_4. The cA_4 second message then activates Csa3a, inducing expression of the acquisition genes, the uptake and expression of new spacers, and their incorporation into both Type-I-A and Type-III surveillance complexes which are then able to cure the infection. In this context, the Type-I-A and Type-III surveillance complexes are not just complementary, they are synergistic, and the integrated response provides a more robust and successful antiviral response than either of the stand-alone systems. In this context, the

multitude of CRISPR loci and cas gene clusters in *S. solfataricus* is far more than a diverse and perhaps redundant collection of various CRISPR-Cas systems, they represent different arms of an integrated class 1 immune system. Csa3 family transcription factors and the cA_4 second messenger lie at the heart of this system, where they coordinate a synergistic, highly orchestrated Class 1 CRISPR-Cas immune response.

4.5 Caveats and Additional Considerations

We cannot rule out the possibility that S. solfataricus Csa3a performs some cA_4-independent functions, and that Csa3 in general may sense other ligands. While Csa3 shows strong association with type I-A systems, it is also occasionally associated with I-B, I-D III-A, III-B and III-D loci [10, 14]. In addition, many genomes encoding Csa3 apparently lack Cas10, and in most Halobacteria and Methanobacteriales, Csa3 is encoded outside the CRISPR-Cas loci [14]. Indeed, Csa3 is even found in organisms apparently lacking CRISPR-Cas [6]. In these cases, it is not clear what the regulatory ligand for Csa3a might be. One exciting possibility, however, is the cyclic-oligonucleotide-based second messengers produced by CD-NTases that are increasingly implicated in a wide range of CBASS antiviral defense systems in bacteria and archaea [55-61]. Csa3 and SAVED domain proteins, for example, may recognize overlapping subsets of these oligonucleotide second messengers.

References

1. Sorek, R., C.M. Lawrence, and B. Wiedenheft, *CRISPR-mediated adaptive immune systems in bacteria and archaea.* Annu Rev Biochem, 2013. **82**: p. 237-66.

2. Hille, F., et al., *The Biology of CRISPR-Cas: Backward and Forward.* Cell, 2018. **172**(6): p. 1239-1259.

3. Charbonneau, A.A., C.C. Gauvin, and C.M. Lawrence, *CRISPR-Cas, A Prokaryotic Adaptive Immune System*, in *Encyclopedia of Biological Chemistry, 3rd Edition*, J. Joseph, Editor. 2021, Oxford: Elsevier. p. 717-741.

4. 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.

5. Sakai, H.D. and N. Kurosawa, *Saccharolobus caldissimus gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of Sulfolobus solfataricus as Saccharolobus solfataricus comb. nov. and Sulfolobus shibatae as Saccharolobus shibatae comb. nov.* Int J Syst Evol Microbiol, 2018. **68**(4): p. 1271-1278.

6. Lintner, N.G., et al., *The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System.* J Mol Biol, 2011. **405**(4): p. 939-955.

7. Lillestøl, R.K., et al., *CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties.* Molecular Microbiology, 2009. **72**(1): p. 259-272.

8. Zink, I.A., E. Wimmer, and C. Schleper, *Heavily Armed Ancestors: CRISPR Immunity and Applications in Archaea with a Comparative Analysis of CRISPR Types in Sulfolobales.* Biomolecules, 2020. **10**(11).

9. 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.

10. Haft, D., et al., *A Guild of 45 CRISPR-Associated (Cas) Protien Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.* Plos Computational Biology, 2005. **1**(6): p. 474-483.

11. Makarova, K.S. and E.V. Koonin, *Annotation and Classification of CRISPR-Cas Systems.* Methods Mol Biol, 2015. **1311**: p. 47-75.

12. Athukoralage, J.S., et al., *Tetramerisation of the CRISPR ring nuclease Crn3/Csx3 facilitates cyclic oligoadenylate cleavage.* Elife, 2020. **9**.

13. Makarova, K.S., et al., *CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems.* Front Genet, 2014. **5**: p. 102.

14. Makarova, K.S., et al., *Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense.* Nucleic Acids Res, 2020. **48**(16): p. 8828-8847.

15. Anantharaman, V., et al., *Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing.* Biol Direct, 2013. **8**: p. 15.

16. Kim, Y.K., Y.G. Kim, and B.H. Oh, *Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of Pyrococcus furiosus.* Proteins, 2013. **81**(2): p. 261-70.

17. Sheppard, N.F., et al., *The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease.* RNA, 2016. **22**(2): p. 216-24.

18. Yan, X., W. Guo, and Y.A. Yuan, *Crystal structures of CRISPR-associated Csx3 reveal a manganese-dependent deadenylation exoribonuclease.* RNA Biol, 2015. **12**(7): p. 749-60.

19. Topuzlu, E. and C.M. Lawrence, *Recognition of a pseudo-symmetric RNA tetranucleotide by Csx3, a new member of the CRISPR associated Rossmann fold superfamily.* RNA Biol, 2016. **13**(2): p. 254-7.

20. Zhu, X. and K. Ye, *Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems.* FEBS Lett, 2012. **586**(6): p. 939-45.

21. Kazlauskiene, M., et al., *A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems.* Science, 2017. **357**(6351): p. 605-609.

22. Niewoehner, O., et al., *Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers.* Nature, 2017. **548**(7669): p. 543-548.

23. Liu, T., et al., *Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition.* Nucleic Acids Res, 2015. **43**(2): p. 1044-55.

24. He, F., et al., *CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b.* Nucleic Acids Res, 2017. **45**(4): p. 1902-1913.

25. Liu, T., et al., *Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus.* Nucleic Acids Res, 2017. **45**(15): p. 8978-8992.

26. Dlakić, M., *HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch.* Bioinformatics, 2009. **25**(23): p. 3071-6.

27. Menon, S.K., et al., *The crystal structure of D212 from sulfolobus spindle-shaped virus ragged hills reveals a new member of the PD-(D/E)XK nuclease superfamily.* J Virol, 2010. **84**(12): p. 5890-7.

28. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein expression and purification, 2005. **41**(1): p. 207.

29. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Analytical Biochemistry, 1976. **72**(1-2): p. 248-254.

30. Brown, S., et al., *Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR-Cas that degrades the second messenger cA4.* J Biol Chem, 2020. **295**(44): p. 14963-14972.

31. Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna, *Mechanism of substrate selection by a highly specific CRISPR endoribonuclease.* RNA (New York, N.Y.), 2012. **18**(4): p. 661-672.

32. Otwinowski, Z. and W. Minor, *[20] Processing of X-ray diffraction data collected in oscillation mode*, in *Methods in Enzymology*. 1997, Elsevier. p. 307-326.

33. Mccoy, A.J., et al., *Phasercrystallographic software.* Journal of Applied Crystallography, 2007. **40**(4): p. 658-674.

34. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D Biological Crystallography, 2010. **66**(2): p. 213-221.

35. Emsley, P., et al., *Features and development ofCoot.* Acta Crystallographica Section D Biological Crystallography, 2010. **66**(4): p. 486-501.

36. Painter, J. and E.A. Merritt, *Optimal description of a protein structure in terms of multiple groups undergoing TLS motion.* Acta Crystallogr D Biol Crystallogr, 2006. **62**(Pt 4): p. 439-50.

37. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallographica Section D Biological Crystallography, 2010. **66**(1): p. 12-21.

38. Delano, W.L., *The PyMOL Molecular Graphics System*. http://www.pymol.org, 2002.

39. Peng, N., et al., *A synthetic arabinose-inducible promoter confers high levels of recombinant protein expression in hyperthermophilic archaeon Sulfolobus islandicus.* Appl Environ Microbiol, 2012. **78**(16): p. 5630-7.

40. Athukoralage, J.S., et al., *Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate.* Nature, 2018. **562**(7726): p. 277-280.

41. Athukoralage, J.S., et al., *An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity.* Nature, 2020. **577**(7791): p. 572-575.

42. Athukoralage, J.S., et al., *The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling.* Elife, 2020. **9**.

43. Athukoralage, J.S., et al., *A Type III CRISPR Ancillary Ribonuclease Degrades Its Cyclic Oligoadenylate Activator.* J Mol Biol, 2019. **431**(15): p. 2894-2899.

44. Garcia-Doval, C., et al., *Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6.* Nat Commun, 2020. **11**(1): p. 1596.

45. Jia, N., et al., *CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity.* Mol Cell, 2019. **75**(5): p. 944-956 e6.

46. Richardson, J.S., et al., *RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution).* RNA, 2008. **14**(3): p. 465-81.

47. Hong, M., et al., *Structure of an OhrR-ohrA operator complex reveals the DNA binding mechanism of the MarR family.* Mol Cell, 2005. **20**(1): p. 131-41.

48. Peng, W., et al., *An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference.* Nucleic Acids Res, 2015. **43**(1): p. 406-17.

49. Baker-Austin, C. and M. Dopson, *Life in acid: pH homeostasis in acidophiles.* Trends Microbiol, 2007. **15**(4): p. 165-71.

50. Moll, R. and G. Schäfer, *Chemiosmotic H+ cycling across the plasma membrane of the thermoacidophilic archaebacterium Sulfolobus acidocaldarius.* FEBS Lett, 1988. **232**(2): p. 359-363.

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

52. Ye, Q., et al., *CRISPR-Associated Factor Csa3b Regulates CRISPR Adaptation and Cmr-Mediated RNA Interference in Sulfolobus islandicus.* Front Microbiol, 2020. **11**: p. 2038.

53. Rouillon, C., et al., *Control of cyclic oligoadenylate synthesis in a type III CRISPR system.* Elife, 2018. **7**.

54. Liu, Z., et al., *A CRISPR-associated factor Csa3a regulates DNA damage repair in Crenarchaeon Sulfolobus islandicus.* Nucleic Acids Res, 2020. **48**(17): p. 9681-9693.

55. Cohen, D., et al., *Cyclic GMP-AMP signalling protects bacteria against viral infection.* Nature, 2019. **574**(7780): p. 691-695.

56. Lowey, B., et al., *CBASS Immunity Uses CARF-Related Effectors to Sense 3'-5'- and 2'-5'-Linked Cyclic Oligonucleotide Signals and Protect Bacteria from Phage Infection.* Cell, 2020. **182**(1): p. 38-49 e17.

57. Lau, R.K., et al., *Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity.* Mol Cell, 2020. **77**(4): p. 723-733 e6.

58. Burroughs, A.M., et al., *Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling.* Nucleic Acids Res, 2015. **43**(22): p. 10633-54.

59. Whiteley, A.T., et al., *Bacterial cGAS-like enzymes synthesize diverse nucleotide signals.* Nature, 2019. **567**(7747): p. 194-199.

60. Shmakov, S.A., et al., *Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis.* Proc Natl Acad Sci U S A, 2018. **115**(23): p. E5307-E5316.

61. White, M.F., *Bacteria SAVED from Viruses.* Cell, 2020. **182**(1): p. 5-6.

CHAPTER FOUR

A CURE FOR HIGH LEVEL GALECTIN EXPRESSION

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Alexander A. Charbonneau

Contributions: Conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing – original draft preparation, writing – review & editing.

Co-Author: Elizabeth J. Reicks

Contributions: Investigation.

Co-Author: John F. Cambria

Contributions: Investigation.

Co-Author: Daria Danley

Contributions: Investigation.

Co-Author: Emmie A. Shockley

Contributions: Investigation, formal analysis.

Co-Author: Ravenor Davion

Contributions: Investigation.

Co-Author: Isabella Salgado

Contributions: Investigation.

Co-Author: Erienne G. Norton

Contributions: Investigation.

Co-Author: Lucy Corbett

Contributions: Investigation.

Co-Author: Lucy E. Hanacek

Contributions: Investigation.

Co-Author: Jordan G. Jensen

Contributions: Investigation, formal analysis.

Co-Author: Marguerite A. Kibodeaux

Contributions: Investigation.

Co-Author: Tess K. Kirkpatrick

Contributions: Investigation.

Co-Author: Keilen M. Rausch

Contributions: Investigation.

Co-Author: Samantha R. Roth

Contributions: Investigation, formal analysis.

Co-Author: Bernadette West

Contributions: Investigation.

Co-Author: Kenai E. Wilson

Contributions: Investigation.

Co-Author: C. Martin Lawrence

Contributions: Conceptualization, methodology, project administration, writing – original draft preparation, writing – review & editing, funding acquisition and resources.

Co-Author: Mary J. Cloninger

Contributions: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing – original draft preparation, writing – review & editing, funding acquisition and resources.

Manuscript Information

Alexander A. Charbonneau, Elizabeth J. Reicks, John F. Cambria, Jacob Inman, Daria Danley, Emmie A. Shockley, Ravenor Davion, Isabella Salgado, Erienne G. Norton†, Lucy Corbett, Lucy E. Hanacek, Jordan G. Jensen, Marguerite A. Kibodeaux, Tess K. Kirkpatrick, Keilen M. Rausch, Samantha R. Roth, Bernadette West, Kenai E. Wilson, C. Martin Lawrence, and Mary J. Cloninger*

Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT, 59717, USA

*Corresponding Author: mcloninger@montana.edu (MJC)

†Current Affiliations: 1) St. Jude Graduate School of Biomedical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA, and 2) Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

Protein Expression and Purification

Status of Manuscript: [Put an x in one of the options below, then delete instruction in brackets]
☐ 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

Elsevier
6 February 2024

Abstract

Galectins are a large and diverse protein family defined by the presence of a carbohydrate recognition domain (CRD) that binds ß-galactosides. They play important roles in early development, tissue regeneration, immune homeostasis, pathogen recognition, and cancer. In many cases, studies that examine galectin biology and the effect of manipulating galectins are aided by, or require the ability to express and purify, specific members of the galectin family. In many cases, *E. coli* is employed as a heterologous expression system, and galectin expression is induced with isopropyl ß-galactoside (IPTG). Here, we show that galectin-3 recognizes IPTG with micromolar affinity and that as IPTG induces expression, newly synthesized galectin can bind and sequester cytosolic IPTG, potentially repressing further expression. To circumvent this putative inhibitory feedback loop, we utilized an autoinduction protocol that lacks IPTG, leading to significantly increased yields of galectin-3. In addition to galectin-3, a member of the chimera subfamily, we also demonstrate high-level expression of galectin-7 and galectin-8, members of the prototype and tandem-repeat subfamilies, respectively. The protocol is thus applicable to galectins in general. Further, much of this work was done within the context of a course-based undergraduate research experience, indicating the ease and reproducibility of the resulting expression and purification protocols.

1. Introduction

Galectins are beta-galactoside binding proteins that play important roles in many intercellular recognition processes including those involved in early development, tissue regeneration, immune homeostasis, pathogen recognition, and cancer progression[1-4]. Fifteen galectins, all including a conserved carbohydrate recognition domain (CRD), have been sorted into three galectin subfamilies[5]. Prototype galectins have two CRDs that form noncovalent homodimers in solution (Fig. 1A); tandem-repeat galectins have an N-terminal CRD, a disordered linker, and a C-terminal CRD (Fig. 1B). Galectin-3, the sole member of the chimera galectin subfamily, has a disordered N-terminal domain (NTD) and a C-terminal CRD (Fig. 1C).

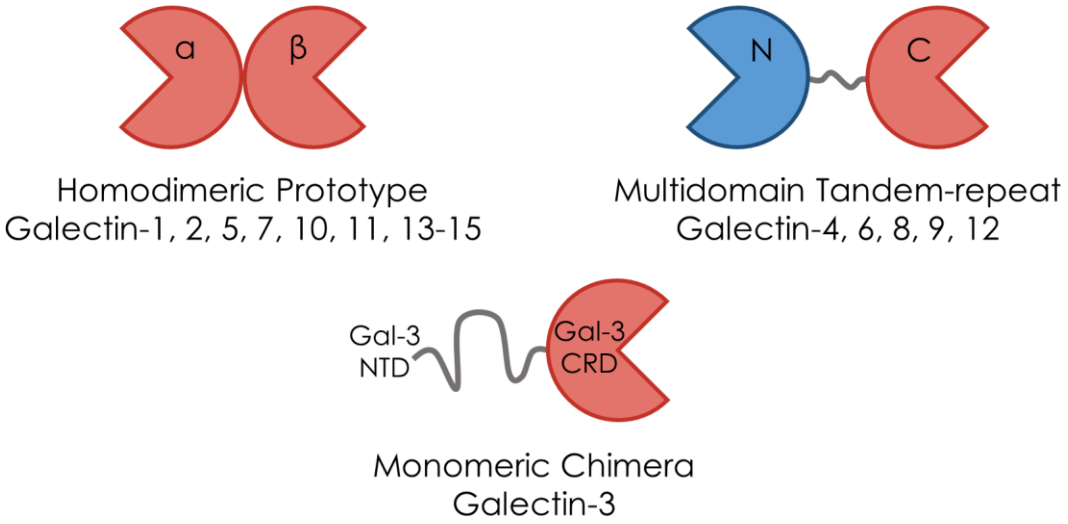


Fig. 1. A schematic representation of the A) prototype, B) tandem-repeat, and C) chimera galectins.

Galectin-3 is reported to aggregate via interactions of its NTD[6], but the need for both dimeric galectins (prototype and tandem-repeat) and a variably oligomeric galectin (chimera) is currently unclear. Histological mapping of the galectins indicates they are broadly distributed in the mammalian body and are co-expressed in many tissues and organs[7], which raises the question of whether galectins from different subfamilies act cooperatively. Indeed, when galectin-1, galectin-3, and galectin-7 were studied as mixtures for *in vitro* cell-based assays, synergistic results were observed[8]. These reports provide a strong impetus for additional studies using galectins from all three subfamilies, both individually and concurrently.

All galectins have a conserved eight-amino acid sequence motif in their carbohydrate-binding pocket[3]. Known native carbohydrate ligands for the galectins have recently been reviewed[4], and although some differences in carbohydrate selectivity have been reported, all galectins bind to beta-galactosides such as lactose[9, 10]. Galectins all have weak monovalent carbohydrate binding interactions, relying on multivalent interactions to increase the specificity and selectivity of their protein/carbohydrate interactions[5, 11-13]. Small molecule inhibitors of galectins have been developed, including thiodigalactosides with aryl groups at the C_3-OH position[14, 15]. The derivatized thiodigalactosides represent an extension of earlier work to develop thiogalactoside derivatives with good galectin-binding activity[16, 17].

Many reports describe the expression of recombinant human galectins in *E. coli[18-21]*. One common and effective method for the purification of these galectins is passage over a lactose affinity column[21, 22]. However, in this case, extensive dialysis to remove lactose is then required. Moreover, for galectin-8 the two domains are often expressed and studied separately [23-25] or a protease-resistant form of galectin-8, in which the disordered linker domain was shortened, is used to avoid proteolytic degradation during the dialysis[26]. While this is an acceptable workaround for some studies, a shorter linker will inevitably change some important properties of galectin-8[27]. For example, the distribution of angles at which galectin-8 presents its two carbohydrate binding domains to a multivalent glycosylated binding partner such as the cell surface or a glycosylated protein in the extracellular matrix would be altered. Insertion of a histidine tag allows for purification using immobilized metal affinity chromatography, and we have found that the presence of the His_6-tag does not change protein activity in binding and aggregation assays[13, 28].

Here, we report enhanced expression and purification of His-tagged galectin-3, galectin-7, and galectin-8. Notably, IPTG is frequently used to induce expression in *E. coli*. However, as shown here, galectins bind IPTG with micromolar affinity, potentially resulting in feedback inhibition of galectin expression and reduced yields. Here, we employ the Studier autoinduction protocol [29] to circumvent this inhibitory feedback loop and demonstrate significantly increased yields for the galectin-3 CRD (Gal-3 CRD). The approach is then extended to other galectin-3

constructs, as well as galectin-7 and galectin-8. We also discuss a course-based undergraduate research experience (CURE) where much of the work was performed. In other words, the "CURE" contributed significantly to a "cure" for high-level galectin expression.

2. Materials and Methods

2.1. Galectin-3 Expression Clones

The expression vector for the full-length human His_6-tagged galectin-3 construct (pEXP14-6xHis-Gal3) was described previously [22]. The galectin-3 carbohydrate recognition domain (Gal-3 CRD, amino acids 112-250) was cloned as previously described[30, 31]. Briefly, fragments of the *Homo sapiens* galectin-3 gene (NM_002306.4) were amplified using polymerase chain reactions (PCR) with the pGEX-6p-Galectin-3 vector as a template. The sequences of the forward and reverse primers are given in Supplementary Table S1. The cloning strategy introduced attB sites, a Shine-Dalgarno sequence, and a minimal N-terminal 6x-His-tag. The amplified products were inserted into pDONR201 (Invitrogen) by site-specific recombination using the Gateway BP clonase reaction following the manufacturer's instructions. Sequence-verified constructs were then transferred into pDEST14 (Invitrogen), again using Gateway site-directed recombination, yielding expression vector pEXP14-6×His-Gal3-CRD.

2.2. Galectin-7 and Galectin-8 Expression Clones

Expression clones for N-terminally His_6-tagged *Homo sapiens* galectin-7 (NM_002307.4, Gal-7) and galectin-8 (NM_201543.4, Gal-8) were ordered from Twist Bioscience. The His-tag added a Met-His_6 sequence at the N-terminus of the native sequence. Clones were based on the pET-29b(+) vector using the NdeI and XhoI restriction sites and were codon optimized for *E. coli* expression. The nucleotide sequence for each clone is provided in Supplementary Fig. S1.

2.3. IPTG Induction

For Gal-3 CRD expression, BL21(DE3)-pRIL *E. coli* (Stratagene) were transformed with pEXP14-6×His-Gal3-CRD. A single colony was used to inoculate 5 mL of LB media (10 g tryptone, 5 g NaCl, and 5 g yeast extract per L) with 100 µg/mL ampicillin and 34 µg/mL chloramphenicol and grown at 37 °C with shaking at 250 rpm for 16 h. Then, 750 mL LB media [32] in a 2.8 L baffled Fernbach flask containing 100 µg/mL ampicillin and 34 µg/mL chloramphenicol was inoculated with 750 µL of the starter culture and grown at 37 °C with shaking at 250 rpm until OD_{600} reached 1.0. IPTG was added to a final concentration of 0 mM, 0.1 mM, 0.5 mM, 1 mM, 2 mM, or 5 mM, and cultures were incubated at 37 °C with shaking at 250 rpm for 18 h. Cells were harvested by centrifugation at 5,000 x g for 20 min, supernatant was discarded, and the pellets were stored at -80 °C.

2.4. Autoinduction

For autoinduction of full-length Gal-3 and Gal-3 CRD, BL21(DE3)-pRIL *E. coli* (Stratagene) were transformed with pEXP14-6×His-Gal3-CRD or pEXP14-6xHis-Gal3 as needed. A single colony was then used to inoculate 5 mL of LB media with 100 µg/mL ampicillin and 34 µg/mL chloramphenicol and grown at 37 °C with shaking at 250 rpm for 4 h. Then, 750 mL ZYP-5052 autoinduction media ([29], see below) in a baffled 2.8 L Fernbach flask containing 100 µg/mL ampicillin and 34 µg/mL chloramphenicol was inoculated with 750 µL of the starter culture and grown at 37 °C with shaking at 250 rpm for 18 h. Cells were harvested by centrifugation at 5,000 x g for 20 min, supernatant was discarded, and the pellets were stored at -80 °C. Autoinduction of galectin-7 (Gal-7) and galectin-8 (Gal-8) were identical to the Gal-3 protocol described above, except 50 µg/mL kanamycin was used instead of ampicillin.

2.5. ZYP-5052 Autoinduction Media

We utilized the autoinduction protocol of Studier[29]. Specifically, 1 L of autoinduction media was composed of 10 g tryptone, 5 g yeast extract, 930 mL H_2O, 1 mL 1 M $MgSO_4$, 100 µL 10,000x trace metal stock solution, and 50 mL 20x NPS. After autoclaving, 20 mL of filter-sterilized 50 x 5052 sugar stock solution was added.

The 10,000x trace metal stock solution was composed of 25 mL 0.1 M $FeCl_3$ in 0.1 M HCl, 0.111 g $CaCl_2$, 0.099 g $MnCl_2 \cdot 4H_2O$, 0.144 g $ZnSO_4 \cdot 7H_2O$, 0.024 g

$CoCl_2 \cdot 6H_2O$, 0.017 g $CuCl_2 \cdot 2H_2O$, 0.024 g $NiCl_2 \cdot 6H_2O$, 0.024 g $Na_2MoO_4 \cdot 2H_2O$, 0.026 g $Na_2SeO_3 \cdot 5H_2O$, 0.006 g H_3BO_4, dissolved in H_2O to a final volume of 50 mL. The 20x NPS stock solution was composed of 33 g $(NH_4)_2SO_4$, 68 g KH_2PO_4, Na_2HPO_4, and dissolved in H_2O to a final volume of 500 mL. The 50x 5052 sugar stock solution was composed of 125 g glycerol, 12.5 g glucose, and 50 g alpha-lactose, dissolved in H_2O to a final volume of 500 mL, and sterilized by passage through a 0.2 μm filter.

2.6. Purification

Cell pellets were thawed and resuspended at 10 mL/g of cell pellet in lysis buffer (400 mM NaCl, 50 mM Tris-Cl, pH 8). PMSF (0.1 mM) was added to the cell suspension, and cells were lysed by passage through a French press (American Instrument Co., Inc., Silver Springs, MD). The lysate was clarified by centrifugation at 30,000 x g for 30 min. The supernatant was then applied to a gravity-flow column containing Ni-NTA agarose (Qiagen). Five flow-through fractions of equal bed volume were collected, and the column was then washed with 10 bed volumes of lysis buffer. Galectins were eluted in 300 mM NaCl, 300 mM imidazole, 50 mM Tris-Cl, pH 8. Protein concentrations were determined by Bradford assay[33] using Protein Assay Reagent (Bio-Rad) and bovine serum albumin as a standard. The purity and mass of each galectin were confirmed by SDS-PAGE. SDS-PAGE gels were imaged with an Invitrogen iBright™CL1500 gel imager. Elution fractions containing galectin were then combined. Following the Ni-NTA purification, Gal-3

CRD and Gal-7, which were generally quite pure, were dialyzed into PBS. Gal-3 and Gal-8 were subjected to size exclusion chromatography on a calibrated Superdex 75 10/300 GL column (Cytiva, Marlborough, MA, USA) with PBS buffer. Following buffer exchange or dialysis into PBS, galectin concentrations were quantified by OD_{280}, the protein was diluted to 1 mg/ml, and stored at -80 ^{o}C in 1 ml aliquots. Extinction coefficients for each construct are given in Supplementary Table S2.

2.7. Analytical Size Exclusion Chromatography

Size exclusion chromatography (SEC), as described above (2.6), was also performed to probe the oligomeric state of each galectin construct. Using PBS, full-length galectin-3 (Gal-3), Gal-3 CRD, and full-length galectin-7 (Gal-7) each showed slower than the expected mobility relative to their monomeric species, signaling an interaction with the matrix (dextran, Supplementary Fig. S2, S3, and S4, respectively). Attempts to disrupt the interaction with increased NaCl concentrations did not resolve this issue. In contrast, Galectin-8 ran as the expected monomer on the SEC column (Supplementary Fig. S5).

2.8. Mass Spectroscopy

Each of the purified proteins was submitted to the IDEA National Resource for Quantitative Proteomics at the University of Arkansas Medical School to confirm the identities of the purified proteins. The protocol for gel-based MS/MS analysis and the data for characterizations of gal-3 CRD, galectin-7, and galectin-8 are provided in the Supplementary Data, Figs. S6, S7, and S8, respectively.

2.9. Fluorescence Lifetime Binding Assays

Fluorescence lifetime binding assays were performed as previously described[28, 34]. Lactose or IPTG were titrated across a 96-well quartz micro-plate (Hellma) for the various galectin constructs at final concentrations ranging from 1.9 to 6.6 μM in PBS, with three technical replicates per plate for each experiment. Wells were excited at 295 nm and time-resolved emission of individual wells followed at 350 nm using a NovaFluor fluorescence lifetime spectrometer (Fluorescence Innovations)[35, 36]. Waveforms were recorded over a total decay time of 128 ns at 2 ps intervals. The fluorescence lifetime data were analyzed as described in the Supplementary Information of Schlick et al. and Bernhard et al.[28, 37]. Briefly, the waveforms collected from each lactose/IPTG concentration were fit to a linear combination of the free (W_f) and complexed (W_c) waveforms, yielding the fraction of protein in complex with ligand at each ligand concentration. The data were then fit to the "one-site specific binding" function in GraphPad Prism ($Y = B_{max}*X/(K_D+X)$; Y = fraction bound; X = ligand concentration).

3. Results

3.1. Low-Level Expression of Gal-3 CRD with IPTG

To produce the *Homo sapiens* Galectin-3 carbohydrate recognition domain (Gal-3 CRD, residues 112-250) to study its interactions with carbohydrates and carbohydrate-functionalized dendrimers[28, 38, 39], we first expressed a minimally, non-cleavable His_6-tagged construct using traditional IPTG induction. Gal-3 CRD

was then purified using immobilized metal affinity chromatography (IMAC) and size exclusion chromatography. Following a previous protocol[40], induction with 0.1 mM IPTG gave 1.5 mg of purified Gal-3 CRD per gram of cell pellet. The purified fractions represented the majority of the Gal-3 CRD, as Gal-3 CRD in the pellet and flow through (FT) fractions appear minimal (Pellet and FT lanes, Fig. 2; Supplementary Table S3).

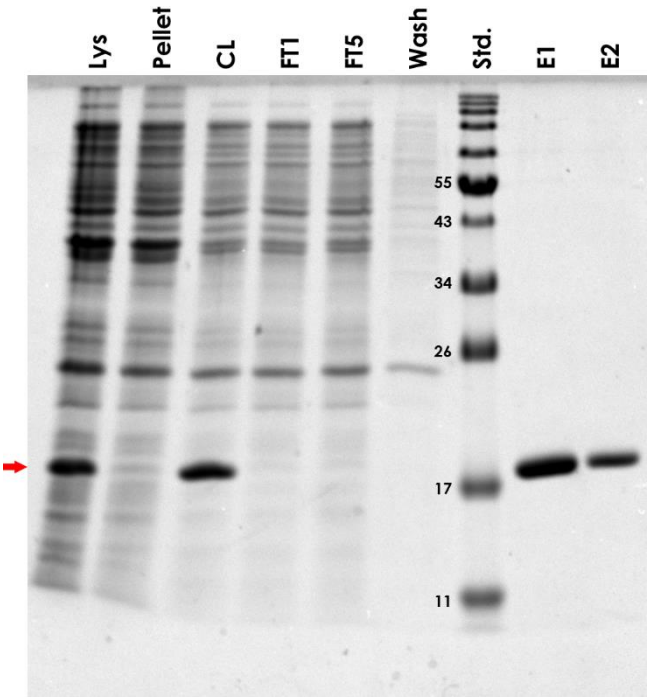


Fig. 2. Gal-3 CRD induction with 0.1 mM IPTG. The lysate (Lys, lane 1) shows a band at approximately 18 kDa (red arrow) consistent with Gal-3 CRD. Gal-3 CRD is also present in the cleared lysate (CL), while flow-through fractions (FT1-2) and the Wash fraction lack quantifiable Gal-3 CRD, signifying Gal-3 CRD is bound to the Ni-NTA resin. Elution fractions (E1-E3) show relatively pure Gal-3 CRD.

To increase the expression levels of Gal-3 CRD, as suggested by Prato et al.[21], we then explored a range of IPTG concentrations, from 0.5 to 2.0 mM. Overall, increased IPTG concentrations gave minimal effect, with yields of 1.35 to

1.79 mg protein/g cell pellet. We then increased the IPTG concentration to 5 mM, resulting in a modest increase in Gal-3 CRD yield of 2.59 mg protein/g cell pellet.

3.2. Basal Gal-3 CRD Expression in the Absence of IPTG

Though our attempts to increase expression levels with increased IPTG gave only a modest increase in yield, to our surprise we found that a control culture lacking IPTG also produced 1.35 mg Gal-3 CRD/g cell pellet (Fig. 3). This suggested much of the Gal-3 CRD expressed in cultures containing ≤ 2 mM IPTG was a result of leaky expression, rather than IPTG induction. Importantly, we also recognized that leaky expression of Gal-3 CRD could result in significant accumulation of Gal-3 CRD before the addition of IPTG. Further, because IPTG is a lactose analogue, we hypothesized that Gal-3 CRD produced by leaky expression might sequester cytosolic IPTG and block induction, resulting in feedback inhibition of Gal-3 CRD expression.

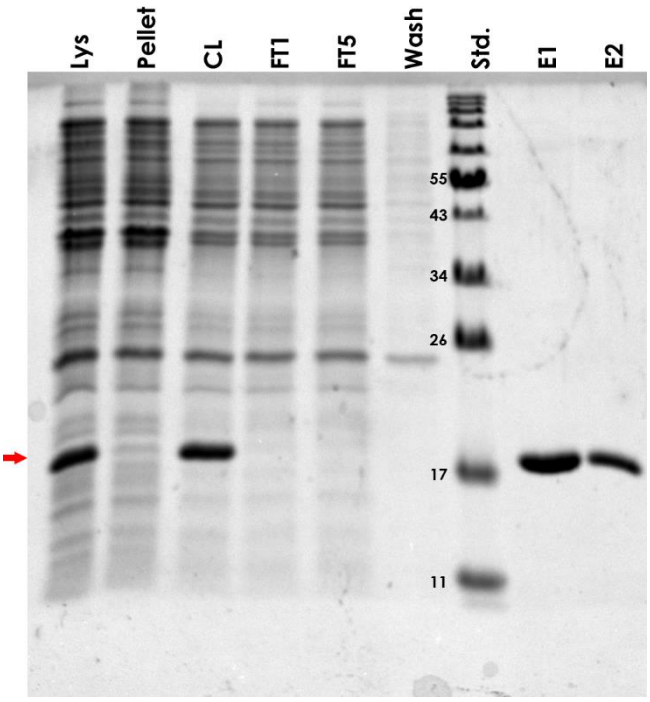


Fig. 3. Leaky expression of Gal-3 CRD. Control cultures grown without IPTG gave 1.35 mg of Gal-3 CRD per gram of cell pellet, nearly equivalent to IPTG-induced cultures. The red arrow (far left) indicates Gal-3 CRD (18 kDa). Lane 1- crude cell lysate (Lys), lane 2 - insoluble pellet, lane 3 - cleared lysate (CL), lanes 4, 5, and 6 - representative flow-through (FT) and Wash fractions, lane 7 - molecular weight standards, lanes 8 and 9 - first and second elutions (E1, E2).

3.3. Gal-3 CRD Binds IPTG

To address the hypothesis that Gal-3 CRD recognizes and sequesters IPTG, and to ensure Gal-3 CRD was properly folded and active for carbohydrate recognition, we measured the ability of Gal-3 CRD to bind both lactose and IPTG. Although many methods have been reported for studying the binding of carbohydrates by galectins[3], time-dependent fluorescence spectroscopy is a facile, highly accurate method for determining binding constants for lectin-carbohydrate interactions[28, 34]. As we previously reported, the fluorescence lifetime of Trp^{181}

can be used to follow ligand (e.g., lactose and IPTG) binding. The mean waveform at each ligand concentration is modeled as a linear combination of the carbohydrate-bound and free waveforms, yielding the fraction bound[28]. As shown in Fig. 4A and B, the time-resolved (sub-nanosecond) fluorescence emission waveforms yielded a K_D of 98 ± 9 μM for Gal-3 CRD binding to lactose (Fig. 4A) and a K_D of 64 ± 5 μM for the interaction between Gal-3 CRD and IPTG (Fig. 4B). Interestingly, the interaction between Gal-3 CRD and IPTG is slightly stronger than the affinity of the Gal-3 CRD for lactose. This is consistent with prior reports by Pieters and coworkers; the replacement of the anomeric oxygen with sulfur affords carbohydrate derivatives that are bound well by galectin CRDs[14, 15]. Similarly, the full galectin-3 construct binds IPTG with a K_D of 42 ± 2 μM (Fig. 4C), which is also slightly higher affinity than our reported value for lactose (K_D = 56 ± 8 μM, also measured in a fluorescence lifetime binding experiment)[28]. When cell density, cell volume, and leaky expression levels of Gal-3 CRD in uninduced cultures are considered[41, 42], we calculate an intracellular Gal-3 CRD concentration of ~ 240 μM, well above the K_D (see "Cellular Concentration of Galectin-3 from Leaky Expression" in Supplementary Data). Thus, the IPTG binding results strengthened the hypothesis that galectins bind and sequester IPTG, in this case, damping induction and expression of Gal-3 CRD in a negative feedback loop.

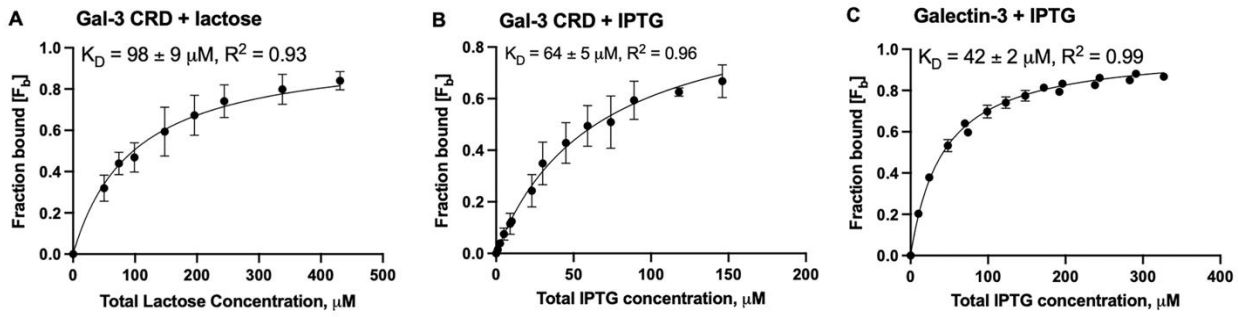


Fig. 4. Affinity of Gallectin-3 for IPTG (A) Gal-3 CRD recognizes lactose with a K_D of 98 ± 9 μM and (B) IPTG with slightly higher affinity, 64 ± 5 μM, while (C) full-length galectin-3 also recognizes IPTG with comparable affinity (K_D= 42 ± 2 μM). Each curve is the mean of at least 3 independent experiments using protein from two different Gal-3 CRD or full-length Galectin-3 preparations. Error bars represent the standard deviation of the mean. The data were fit to the one-site single binding function in GraphPad Prism.

3.4. Autoinduction Produces High Gal-3 CRD Yields

After identifying the potential for Gal-3 CRD to feedback inhibit its induction with IPTG, we looked to autoinduction[29], a tool our lab has used extensively[31, 43, 44]. Similar to IPTG, autoinduction also utilizes the T7 expression system. However, rather than IPTG, autoinduction instead utilizes specific ratios of carbon sources, in this case glycerol, glucose, and lactose, to tune induction. Importantly, this allowed us to utilize our existing vectors. We found that autoinduction reliably produced 10.8 mg ± 1.0 mg Gal-3 CRD per gram of cell pellet (Fig. 5, Supplementary Data Table S4), a more than 7-fold increase relative to 1 mM IPTG, and nearly 4 times more than 5.0 mM IPTG (Fig. 6, Supplementary Data Table S5).

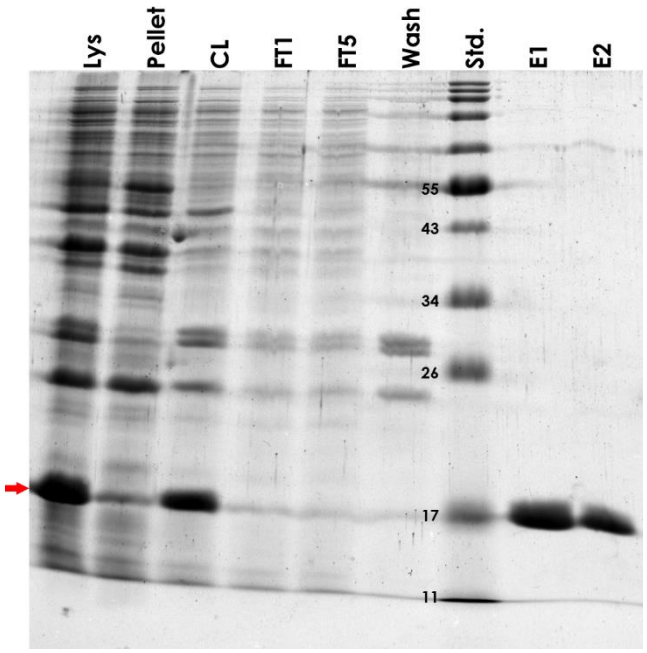


Fig. 5. Autoinduction increases Gal-3 CRD expression. The red arrow (far left) indicates Gal-3 CRD (18 kDa). Lane 1- crude cell lysate (Lys), lane 2 - insoluble pellet, lane 3 - cleared lysate (CL), lanes 4, 5 and 6 - representative flow-through (FT) and wash fractions, lane 7 - molecular weight standards, lanes 8 and 9 - first and second elutions (E1, E2).

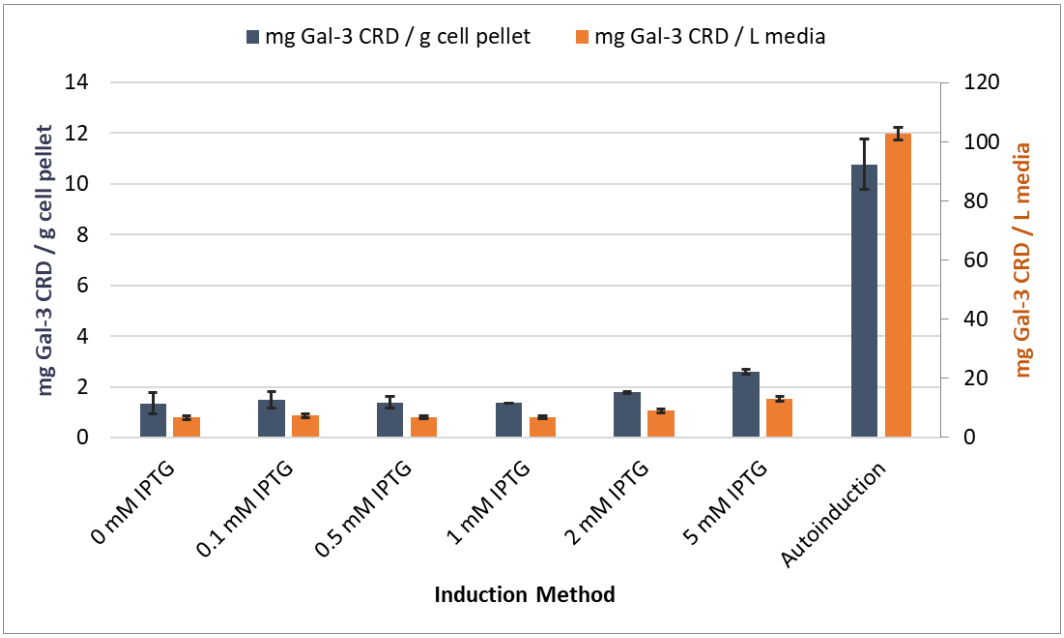


Fig. 6. Autoinduction outperforms IPTG. Autoinduction results in greater yields than IPTG induction, in terms of both mg protein/g cell pellet (blue bars) and mg protein/L of culture (orange bars). We consistently purified about 7 times more Gal-3 CRD per gram of cell pellet using autoinduction compared to induction with 0.1 to 1 mM IPTG.

3.5. Autoinduction, Expression, and Purification of Additional Galectins

Following successful expression of Gal-3 CRD with autoinduction media, we sought to express full-length constructs for a representative member of each of the three classes of galectins: galectin-3 (chimera), galectin-7 (homodimer) and galectin-8 (prototype) (Fig. 1). Using the same autoinduction protocol, expression and purification of full-length galectin-3 (Gal-3) resulted in a yield of 2.5 mg/g cell pellet or 25.0 mg/L culture (10 g cell pellet/L culture, lane 2 of the gel shown in Fig. 7). High yields of galectin-7 and galectin-8 were also achieved. Galectin-7 (Gal-7, Fig. 7) gave yields of 9.0 mg/g cell pellet or 90 mg/L culture (10 g cell pellet/L culture).

Galectin-8 (Gal-8, Fig. 7) was expressed at slightly lower levels with a yield of 3.0 mg/g cell pellet or 30 mg/L culture (10 g cell pellet/L culture).

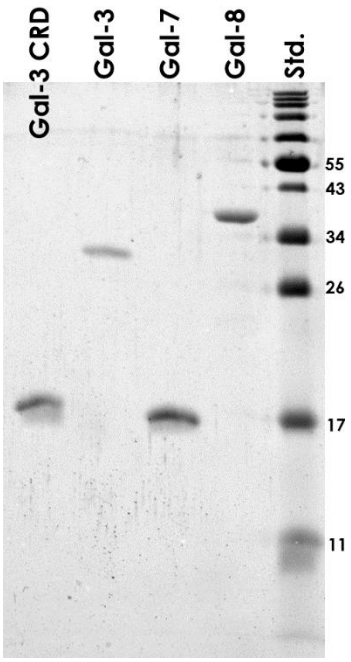


Fig. 7. Autoinduction efficiently produces each class of human galectin. Lane 1 – Gal-3 CRD, lane 2 – full-length Gal-3 (chimera), lane 3 – Gal-7 (homodimer), lane 4 – Gal-8 (prototype). *Homo sapiens* Gal-3 CRD runs at 18 kDa, while full-length Gal-3, Gal-7, and Gal-8 run at 32 kDa, 17 kDa and 38 kDa, respectively.

4. Discussion

4.1. Comparison ofAautoinduction and IPTG-Induction for Galectin Expressions

In classic work, Jacques Monod demonstrated diauxic growth in mixtures of glucose and lactose, where glucose, as the preferred carbon source, is metabolized first[45, 46]. Glucose also inhibits uptake and catabolism of glycerol[47]. In this light, the Studier autoinduction protocol utilizes N-Z amine (amino acids), yeast extract and a mixture of glucose (0.05%), lactose (0.2%) and glycerol 0.5%)[48].

When cultures are first inoculated in the presence of 0.05% glucose, the uptake of lactose and glycerol are inhibited. However, as glucose and amino acids are exhausted, cAMP levels rise and engage the catabolite activator protein (CAP)[49]. Carbon catabolite repression is thus relieved (CCR, reviewed in[46]), and lactose and glycerol are then co-metabolized [47, 48]. Importantly, in *E. coli* strains harboring prophage lambda DE3, the increased lactose levels drive expression of T7 RNA polymerase, which is under control of the lacUV5 promoter (P_{lacUV5})[50, 51]. T7 RNA polymerase, in-turn, drives high-level transcription of a target gene under control of the T7 promoter, resulting in robust overexpression of the gene of interest [50, 51].

With traditional IPTG induction protocols using LB media, IPTG concentrations are low (~ 1 mM), and thus potentially susceptible to sequestration by overexpressed galectin (Fig. 8). In contrast, with autoinduction media, lactose concentrations are significantly higher (0.2%, 6 mM). And though lactose concentrations begin to fall as glucose is exhausted, this happens relatively slowly due to the presence of 54 mM glycerol as the major carbon source [48]. For this reason, free intracellular lactose is sustained, providing the requisite levels of T7 RNA polymerase needed for high level transcription, as well as de-repression of the lac operator present in expression plasmids like pET28.

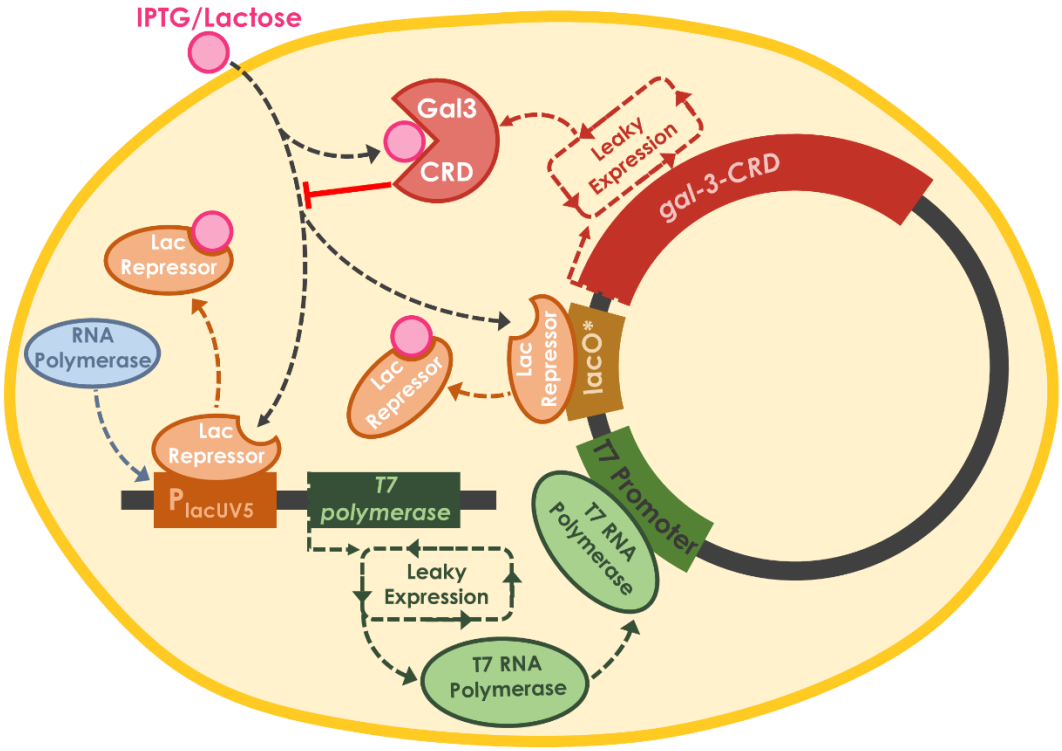


Fig. 8. Feedback inhibition. *E. coli* BL21 (DE3) carries a chromosomal copy of T7 RNA polymerase under control of the P_{lacUV5} promoter, such that expression of T7 RNA polymerase is inhibited by lac repressor (LacI). Upon addition of IPTG, transcriptional repression is relieved, driving expression of T7 RNA polymerase, which then binds the T7 promoter in various expression plasmids, inducing expression of the target gene. When the target gene is a galectin such as Gal-3 CRD, accumulating galectin levels may in-turn bind and sequester intracellular IPTG. Gal-3 CRD potentially outcompetes and deprives lac repressor of IPTG, resulting in repression of both T7 RNA polymerase and *gal-3 crd*. Gal-3 CRD may thus feedback inhibits its own production. Even in the absence of IPTG, low level expression of T7 RNA polymerase from the P_{lacUV5} promoter generally occurs, which may drive leaky expression of the target gene[51]. This is especially true in vectors like pDEST14 that lack secondary induction control, such as the lac operator (*). In the case of Gal-3 CRD, leaky expression levels appear to be high enough to sequester intracellular IPTG, inhibiting further production of Gal-3 CRD.

Autoinduction media (ZYP-5052)) is a rich media, and from this point of view, is expected to provide increased cell mass, and thus greater yields in terms of mg protein per liter of culture[48]. In our work, IPTG-induced LB media produced 5.0 g

± 0.3 g of cells per liter of media, whereas cell pellets derived from autoinduction were 9.5 g ± 0.2 g/L, roughly a two-fold difference. It is for this reason that we have instead characterized the increased yield with respect to 1 gram of cell pellet. As shown in Fig. 6, comparing mg of Gal-3 CRD purified from a gram of cell pellet (blue bars in Fig. 6), autoinduction increased yields by 7-fold compared to IPTG. Nevertheless, when yields are standardized to culture volume (orange bars in Fig. 6), the efficiency of autoinduction increases to ~14-fold higher than IPTG-induction. In addition, because autoinduction eliminates the need for monitoring of optical density prior to addition of IPTG, autoinduction is a facile method for protein expression, one that worked extremely well within the context of our course based undergraduate research experience.

4.2. A "CURE" for Galectin Expression, Purification, and Characterization

4.2.1. Overview and Comparison of CUREs and UREs. Much of this work was conducted within the context of a course-based undergraduate research experience, or CURE. CUREs are an increasingly popular strategy for providing research experiences to undergraduates [52-54]. When compared to more traditional undergraduate research experiences (UREs) within the faculty instructor's research program, CUREs have the advantage that supervision by one faculty instructor (or one team of faculty members and graduate students) can include a large number of undergraduate researchers. Compared to the small, or even 1:1, student-to-faculty instructor ratio employed in traditional UREs, the CURE model engages a

significantly larger number of undergraduates in research activities. In addition, students ranging from the introductory through the advanced undergraduate levels can be included more seamlessly in a CURE, while UREs often include primarily senior-level undergraduate researchers. Finally, CUREs may be a more inclusive entry point to scientific research than UREs, and the implementation of CUREs at the introductory level may therefore be a way to improve the diversity of the scientific community [55-57].

Auchincloss and colleagues were among the first to use the CURE terminology [52]. Their description of the curricular components of a CURE includes the following five areas. Discovery includes the achievement of new knowledge or insight via student work. Relevance requires that student work have an impact beyond the classroom. Collaboration requires that students work in groups. Iteration enables student participants to conduct similar experiments repetitively for improvement of their technical skills and scientific knowledge. Iteration also enables student participants to build upon knowledge obtained by other CURE students. Finally, the CURE includes science practices such as developing a hypothesis/research question, designing methodology to address the hypothesis, reviewing primary literature, and disseminating research results [52, 53]. In the next section, we describe how our CURE was structured for the expression and study of galectins.

In their recent review of CUREs across the chemistry curriculum, Watts and Rodriguez stated that biochemistry CUREs "typically involved studying

relationships among sequence, structure, and function for proteins or enzymes that remain understudied in the scientific literature" [54]. Indeed, recent examples such as the CURE reported by Battin and coworkers suggest that this is a rich subject area for CUREs [58]. With this in mind, we hypothesized that expression of His_6-tagged galectins from rich autoinduction media would not only improve the yields and purities of galectins. In addition, we hypothesized that a CURE would be an effective way to obtain and study galectins. How this CURE was structured, and how it could be translated for other biochemistry CUREs, is presented in subsection 4.2.2.

4.2.2. A CURE for Galectin Expression, Purification, and Characterization.

Fifteen of the coauthors on this manuscript are undergraduates who performed the described galectin research as part of the CURE. Two of the coauthors are high school students, the first author is the graduate teaching assistant for the CURE, and the two final authors are the faculty who supervised the CURE for galectin expression and purification.

All biochemistry majors at Montana State University are required to take a minimum of six credits of undergraduate research during their program of study, which until now has been accomplished via UREs in the research programs of faculty members in the Department of Chemistry and Biochemistry. To address the need for additional UREs within our rapidly growing biochemistry major, Charbonneau, Lawrence, and Cloninger developed a CURE. Students registered for

2 or 3 credits for undergraduate research and spent at least six (2 credits) or nine (3 credits) hours in lab each week. Students were allowed to sign up for blocks of time (2 hours minimum) within an available 28 hours: Monday, Tuesday, and Wednesday 9-5, and Thursday 1-5. Students were required to check in at the beginning of their lab time with one of the three CURE instructors, and the timing was arranged such that students overlapped all day long to ensure that communication worked well, students were ensured of a safe and productive learning environment, and protein purifications were completed.

Often, at least six different students would work on various stages of one galectin purification and characterization protocol. For example, morning students would make buffers, thaw and resuspend cell pellets, lyse cells (French Press), and perform gravity flow Ni-NTA chromatography. Afternoon students would perform Bradford assays, SDS-PAGE, dialysis or size exclusion chromatography. Day two students would complete size exclusion chromatography and prepare purified galectins for storage at -80 ^{o}C. Over the course of 3-4 weeks, each student was taught all aspects of the expression, purification, and characterization pipeline, while the lysis, wash, and elution buffers were concomitantly optimized by teams of students using buffers with varied pH and salt concentrations. These teams worked side-by-side on simultaneous purifications so that protocols could be efficiently optimized.

During the first third of the first semester, in lab practical undergraduates were taught to perform the protocols and techniques required for sections 2.4-2.7

above. During the second third of the semester, students repeated these protocols to optimize conditions and to ensure that all students learned all aspects of the expression, purification, and characterization protocols. During the final third of the semester, purifications to obtain needed galectins continued. In addition, for assessment, each student chose one day to independently perform a purification and characterization protocol. Galectin-3, Gal-3 CRD, and galectin-7 were obtained during the first semester CURE.

At the start of the second semester, two new undergraduate students were added to the CURE cohort, and two high school students and one summer research undergraduate were added to the cohort during the summer. The continuing undergraduates taught the new students how to perform galectin purifications so that additional galectin-3, Gal-3 CRD, and galectin-8 could be prepared. In the summer, the continuing undergraduates taught the high school students how to purify galectins, how to cast gels, how to run Bradford assays, and how to perform time-dependent fluorescence binding assays as described in section 2.9.

During the CURE, students attended weekly classroom sessions. Topics included biosafety training from MSU's Biosafety Officer, introductory information on galectins, and discussion of the theoretical and practical aspects of electronic records, pipetting, buffers, column chromatography, Bradford assays, SDS-PAGE, basic microbiology skills and techniques, plasmids and transcriptional regulation. The remainder of the classroom sessions were focused on teaching students to effectively communicate their scientific results. Students prepared 10–15-minute

presentations of the research they had performed during the prior two weeks. They presented their proposed experiments for the next two weeks, and the cohort spent time troubleshooting, identifying areas for improvement in the experimental protocols, and asking/answering questions about the projects.

Assessments for this course included evaluation of student presentations, including the students' abilities to answer questions from cohort instructors. The individualized, end-of-semester galectin purification and characterizations were also used for evaluation. Assessment of persistence by CURE participants to obtain their undergraduate degrees as well as persistence in biochemistry (or related) research will be performed by surveying these students annually for the next five years.

The CURE instructors and students both noted several important advantages afforded by the CURE structure. First, the CURE students were able to appreciably advance this area of research. When undergraduates are working only 6-10 hours per week on a non-collaborative project in a traditional URE, similar progress can occur but is much less assured. CURE students also appreciated the classroom instruction, the opportunity to learn how to structure PowerPoint slides for scientific presentations, and the camaraderie that they established within their CURE cohort. They felt that they learned more by both performing research and communicating regularly as they transitioned from one researcher to another throughout the day.

One key difference between this CURE and the majority of CUREs is that students were not restricted to one or two lab sessions per week. Moving forward, this CURE cohort will transition to galectin-glycopolymer studies and cell-based assays. Additional CUREs focused on protein expression, purification, and characterization could be readily accomplished with a graduate teaching assistant (GTA) and 2-3 advanced undergraduate CURE participants. We find that this model works well with up to 15 students. This CURE model moreover works best when a graduate teaching assistant is embedded within the CURE laboratory, and participating faculty are in close proximity, as this enables the GTA and faculty and to provide almost constant oversight of the CURE while still accomplishing their standard duties.

As demonstrated in this manuscript, this CURE accomplished nearly all the major CURE components described above [52-54]. Students experienced “discovery” when they successfully used autoinduction to prepare galectin-3, Gal-3 CRD, galectin-7, and galectin-8. The students grasped the “relevance” of their CURE research as they began to perform studies using the galectins they isolated to elucidate the role of galectins during many stages of cancer progression. The students saw daily how effectively “collaboration” could be used to move their research projects forward, even though they had to leave the lab to attend their classes. The students performed many “iterations” of their purification protocols to optimize the lysis, wash, and elution buffer salt concentrations, to determine the best cell lysis protocol, etc. Regarding “science practices”, the primary hypothesis

driving the work (e.g., autoinduction should increase yields of galectins) was established by the instructors prior to the start of the CURE. However, smaller hypotheses such as which salt concentration would be best in the wash buffer were developed and tested routinely by the undergraduates. In addition, CURE students routinely helped to "design methodology" during classroom discussions sessions and were provided with relevant papers and encouraged to "review primary literature" prior to these sessions. The CURE is also "disseminating results", not only with this manuscript but in poster presentations at the MSU Undergraduate Research Celebrations and the American Chemical Society Northwest Regional Meeting (June, 2023, Bozeman, MT).

5. Conclusion

The autoinduction protocol described here was used to produce His_6-tagged Gal-3 CRD, as well as full length Gal-3, Gal-7, and Gal-8, each with high yield and purity. The protocol also eliminates the need to monitor *E. coli* growth, making autoinduction a facile method for novice researchers, and thus ideal for a course-based undergraduate research experiences (CURE). Time-dependent fluorescence shows that galectins efficiently bind IPTG. Galectins can thus feedback inhibit their own expression when IPTG is used. Autoinduction overcomes this barrier to high level galectin expression.

References

1. Marino, K.V., et al., *Targeting galectin-driven regulatory circuits in cancer and fibrosis.* Nature Reviews Drug Discovery, 2023. **22**(4): p. 295-316.

2. Navarro, P., et al., *Impact of Galectins in Resistance to Anticancer Therapies.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2020.

3. Arthur, C.M., et al., *Evolving Mechanistic Insights into Galectin Functions*, in *Galectins: Methods and Protocols*, S.R. Stowell and R.D. Cummings, Editors. 2015. p. 1-35.

4. Troncoso, M.F., et al., *The Universe of Galectin-Binding Partners and Their Functions in Health and Disease.* The Journal of biological chemistry, 2023: p. 105400-105400.

5. Caballero, G.G., et al., *How galectins have become multifunctional proteins.* Histology and Histopathology, 2020. **35**(6): p. 509-539.

6. Sciacchitano, S., et al., *Galectin-3: One Molecule for an Alphabet of Diseases, from A to Z.* International Journal of Molecular Sciences, 2018. **19**(2): p. 59.

7. Nio-Kobayashi, J., *Histological Mapping and Subtype-Specific Functions of Galectins in Health and Disease.* Trends in Glycoscience and Glycotechnology, 2018. **30**(172): p. SE89-SE96.

8. Dings, R.P.M., et al., *Simulating cellular galectin networks by mixing galectins in vitro reveals synergistic activity.* Biochemistry and biophysics reports, 2021. **28**: p. 101116-101116.

9. Stowell, S.R. and R.D. Cummings, *Galectins: Methods and Protocols*. 2015, New York: Springer Science.

10. Cummings, R.D. and F.T. Liu, *Galectins*, in *Essentials of Glycobiology. 2nd ed.*, A. Varki, et al., Editors. 2009, Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY).

11. Gabius, H.J., et al., *What is the Sugar Code?* Chembiochem.

12. Di Lella, S., et al., *When Galectins Recognize Glycans: From Biochemistry to Physiology and Back Again.* Biochemistry, 2011. **50**(37): p. 7842-7857.

13. Bernhard, S.P., et al., *Protein aggregation nucleated by functionalized dendritic polyglycerols.* Polymer Chemistry, 2020. **11**(23): p. 3849-3862.

14. van Hattum, H., et al., *Tuning the Preference of Thiodigalactoside- and Lactosamine-Based Ligands to Galectin-3 over Galectin-1.* Journal of Medicinal Chemistry, 2013. **56**(3): p. 1350-1354.

15. van Scherpenzee, M., et al., *Synthesis and Evaluation of New Thiodigalactoside-Based Chemical Probes to Label Galectin-3.* Chembiochem, 2009. **10**(10): p. 1724-1733.

16. Giguère, D., et al., *Aryl O- and S-galactosides and lactosides as specific inhibitors of human galectins-1 and -3:: Role of electrostatic potential at O-3.* Bioorganic & Medicinal Chemistry Letters, 2006. **16**(6): p. 1668-1672.

17. Cumpstey, I., et al., *Synthesis of a phenyl thio-β-D-galactopyranoside library from 1,5-difluoro-2,4-dinitrobenzene:: discovery of effcient and selective monosaccharide inhibitors of galectin-7.* Organic & Biomolecular Chemistry, 2005. **3**(10): p. 1922-1932.

18. Dam, T.K., et al., *Galectins bind to the multivalent glycoprotein asialofetuin with enhanced affinities and a gradient of decreasing binding constants.* Biochemistry, 2005. **44**(37): p. 12564-12571.

19. Poland, P.A., C.L. Kinlough, and R.P. Hughey, *Cloning, Expression, and Purification of Galectins for In Vitro Studies.* Galectins: Methods and Protocols, 2015. **1207**: p. 37-49.

20. Poland, P.A., C.L. Kinlough, and R.P. Hughey, *Cloning, Expression, and Purification of Galectins for In Vitro Studies.* Methods in molecular biology (Clifton, N.J.), 2022. **2442**: p. 41-54.

21. Prato, C.A., et al., *Purification of Recombinant Galectins Expressed in Bacteria.* Star Protocols, 2020. **1**(3).

22. Dey, C., P. Palm, and L. Elling, *Characterization of Galectin Fusion Proteins with Glycoprotein Affinity Columns and Binding Assays.* Molecules, 2023. **28**(3).

23. Bohari, M.H., et al., *Structure-Based Design of a Monosaccharide Ligand Targeting Galectin-8.* Chemmedchem, 2018. **13**(16): p. 1664-1672.

24. Chien, C.T.H., et al., *Lactose Binding Induces Opposing Dynamics Changes in Human Galectins Revealed by NMR-Based Hydrogen-Deuterium Exchange.* Molecules, 2017. **22**(8).

25. Zhang, H., et al., *Discovery of N-Arylsulfonyl-Indole-2-Carboxamide Derivatives as Galectin-3 and Galectin-8 C-Terminal Domain Inhibitors.* Acs Medicinal Chemistry Letters, 2023. **14**(9): p. 1257-1265.

26. Yoshida, H., et al., *X-ray structure of a protease-resistant mutant form of human galectin-8 with two carbohydrate recognition domains.* Febs Journal, 2012. **279**(20): p. 3937-3951.

27. Si, Y., et al., *Linker remodels human Galectin-8 structure and regulates its hemagglutination and pro-apoptotic activity.* International Journal of Biological Macromolecules, 2023. **245**.

28. Bernhard, S.P., et al., *Time-Dependent Fluorescence Spectroscopy to Quantify Complex Binding Interactions.* ACS Omega, 2020. **5**(45): p. 29017-29024.

29. Studier, F.W., *Protein production by auto-induction in high-density shaking cultures.* Protein Expression and Purification, 2005. **41**(1): p. 207-234.

30. Menon, S.K., et al., *The Crystal Structure of D212 from Sulfolobus Spindle-Shaped Virus Ragged Hills Reveals a New Member of the PD-(D/E)XK Nuclease Superfamily.* Journal of Virology, 2010. **84**(12): p. 5890-5897.

31. Lintner, N.G., et al., *The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system.* J Mol Biol, 2011. **405**(4): p. 939-55.

32. Bertani, G., *STUDIES ON LYSOGENESIS I.* Journal of Bacteriology, 1951. **62**(3): p. 293-300.

33. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Analytical Biochemistry, 1976. **72**(1-2): p. 248-254.

34. Brown, S., et al., *Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR-Cas that degrades the second messenger cA(4).* J Biol Chem, 2020. **295**(44): p. 14963-14972.

35. Petersen, K.J., et al., *Fluorescence lifetime plate reader: Resolution and precision meet high-throughput.* Review of Scientific Instruments, 2014. **85**(11): p. 113101.

36. Muretta, J.M., et al., *High-performance time-resolved fluorescence by direct waveform recording.* Review of Scientific Instruments, 2010. **81**(10): p. 103101.

37. Schlick, K.H., et al., *Characterization of Protein Aggregation via Intrinsic Fluorescence Lifetime.* Journal of the American Chemical Society, 2009. **131**(46): p. 16608-16609.

38. Ennist, J.H., et al., *Chemoenzymatic Synthesis of Galectin Binding Glycopolymers.* Bioconjugate Chemistry, 2018. **29**(12): p. 4030-4039.

39. Goodman, C.K., et al., *Multivalent scaffolds induce galectin-3 aggregation into nanoparticles.* Beilstein Journal of Organic Chemistry, 2014. **10**: p. 1570-1577.

40. Larson, E.T., et al., *Structure of A197 from Sulfolobus Turreted Icosahedral Virus: a Crenarchaeal Viral Glycosyltransferase Exhibiting the GT-A Fold.* Journal of Virology, 2006. **80**(15): p. 7636-7644.

41. Kubitschek, H.E. and J.A. Friske, *DETERMINATION OF BACTERIAL-CELL VOLUME WITH THE COULTER-COUNTER.* Journal of Bacteriology, 1986. **168**(3): p. 1466-1467.

42. Myers, J.A., B.S. Curtis, and W.R. Curtis, *Improving accuracy of cell and chromophore concentration measurements using optical density.* Bmc Biophysics, 2013. **6**.

43. Brown, S., et al., *Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR–Cas that degrades the second messenger cA4.* Journal of Biological Chemistry, 2020. **295**(44): p. 14963-14972.

44. Charbonneau, A.A., et al., *Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response in Saccharolobus solfataricus.* Biomolecules, 2021. **11**(12): p. 1852.

45. Monod, J., *The growth of bacterial cultures.* Annual review of microbiology, 1949. **3**(1): p. 371-349.

46. Stulke, J. and W. Hillen, *Carbon catabolite repression in bacteria.* Curr Opin Microbiol, 1999. **2**(2): p. 195-201.

47. Bettenbrock, K., et al., *A quantitative approach to catabolite repression in Escherichia coli.* J Biol Chem, 2006. **281**(5): p. 2578-84.

48. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. **41**(1): p. 207-34.

49. Voet, D. and J.G. Voet, *Biochemistry*. 4th ed. 2011, Hoboken, NJ: John Wiley & Sons. xxv, 1428, 53 p.

50. Studier, F.W. and B.A. Moffatt, *Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes.* J Mol Biol, 1986. **189**(1): p. 113-30.

51. Studier, F.W., et al., *Use of T7 RNA polymerase to direct expression of cloned genes.* Methods Enzymol, 1990. **185**: p. 60-89.

52. Auchincloss, L.C., et al., *Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report.* Cbe-Life Sciences Education, 2014. **13**(1): p. 29-40.

53. Buchanan, A.J. and G.R. Fisher, *Current Status and Implementation of Science Practices in Course-Based Undergraduate Research Experiences (CUREs): A Systematic Literature Review.* Cbe-Life Sciences Education, 2022. **21**(4).

54. Watts, F.M. and J.-M.G. Rodriguez, *A Review of Course-Based Undergraduate Research Experiences in Chemistry.* Journal of Chemical Education, 2023. **100**(9): p. 3261-3275.

55. Bangera, G. and S.E. Brownell, *Course-Based Undergraduate Research Experiences Can Make Scientific Research More Inclusive.* Cbe-Life Sciences Education, 2014. **13**(4): p. 602-606.

56. Rodenbusch, S.E., et al., *Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees.* Cbe-Life Sciences Education, 2016. **15**(2).

57. Estrada, M., et al., *Improving Underrepresented Minority Student Persistence in STEM.* Cbe-Life Sciences Education, 2016. **15**(3).

58. Groover, S., J. Legleiter, and E.E. Battin, *Huntingtin Protein Purification and Experimentation: An Innovative Undergraduate Research-Based Biochemistry Experiment.* Journal of Chemical Education, 2021. **98**(9): p. 3011-3018.

CHAPTER FIVE

CSX3 IS A CYCLIC OLIGONUCLEOTIDE PHOSPHODIESTERASE ASSOCIATED WITH TYPE-III CRISPR-CAS THAT DEGRADES THE SECOND MESSENGER CA_4

Contribution of Authors and Co-Authors

Manuscript in Appendix C

Author: Sharidan Brown

Contributions: Data curation, formal analysis, investigation.

Co-Author: Colin C. Gauvin

Contributions: Formal analysis, investigation, writing-review and editing, writing-original draft.

Co-Author: Alexander A. Charbonneau

Contributions: Formal analysis, investigation, writing-review and editing, writing-original draft.

Co-Author: Nathaniel Burman

Contributions: Investigation, writing-review and editing, writing-original draft.

Co-Author: C. Martin Lawrence

Contributions: Data curation, formal analysis, investigation, writing-review and editing, writing-original draft, conceptualization, supervision, funding acquisition, methodology, project administration.

Manuscript Information

Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type-III CRISPR-Cas that degrades the second messenger cA_4

Sharidan Brown[1], Colin C. Gauvin[1,2], Alexander A. Charbonneau[1,2], Nathaniel Burman[1] and C. Martin Lawrence[1,2*]

[1]Department of Chemistry and Biochemistry and the [2]Thermal Biology Institute

Montana State University, Bozeman, MT 59717, USA

Journal of Biological Chemistry
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

Elsevier
30 October 2020
Journal of Biological Chemistry 2020, Volume 295, Issue 44
https://doi.org/10.1074/jbc.RA120.014099

Abstract

Cas10 is the signature gene for type-III CRISPR-Cas surveillance complexes. Unlike type-I and type-II systems, type-III systems do not require a PAM, and target nascent RNA associated with transcriptionally active DNA. Further, target RNA recognition activates the cyclase domain of Cas10, resulting in the synthesis of cyclic oligoadenylate (cOA) second messengers. These second messengers are recognized by ancillary Cas proteins harboring CARF domains, and regulate the activities of these proteins in response to invading nucleic acid. Csx3 is a distant member of the CARF domain superfamily previously characterized as a Mn2+ dependent deadenylation exoribonuclease. However, its specific role in CRISPR-Cas defense remains to be determined. Here we show Csx3 is strongly associated with type-III systems and that Csx3 binds cyclic tetra-adenylate (cA4) second messenger with high affinity. Further, Csx3 harbors cyclic oligonucleotide phosphodiesterase activity that quickly degrades this cA4 signal. In addition, structural analysis identifies core elements that define the CARF domain fold, and the mechanistic basis for ring nuclease activity are discussed. Overall, the work suggests Csx3 functions within CRISPR-Cas as a counterbalance to Cas10 to regulate the duration and amplitude of the cA4 signal, providing an off-ramp from the programmed cell death pathway in cells that successfully cure viral infection.

Introduction

CRISPR-Cas is a prokaryotic adaptive immune system that is now well known for its repurposed uses in genome editing. Within the prokaryotic domains of life where these systems evolved, a multitude of various classes, types and subtypes are now recognized. In each case, however, the basic mechanism of adaptive immunity is generally described in three distinct stages. Briefly, in stage 1 (adaptation) short fragments of invading DNA known as spacers are inserted into the CRISPR locus. During stage 2 (crRNA biogenesis) the CRISPR locus is transcribed and cleaved into short CRISPR RNAs (crRNAs), and in stage 3 (target interference) crRNA is assembled into nucleoprotein surveillance complexes that survey the cell for complementary DNA (type-I, type-II) or RNA (type-III). Recognition of target DNA or RNA then activates nucleases that destroy the invading nucleic acid, providing immunity. Further, in type-III systems, which recognize RNA associated with transcriptionally active DNA, not only is the RNA degraded, but the DNA as well [1, 2].

This process is wholly dependent on a wealth of CRISPR associated (Cas) protein machinery. Cas1, Cas2 and Cas4 are generally involved in adaptation (spacer acquisition), Cas6 frequently in crRNA biogenesis, and diverse groups of type and subtype dependent proteins are utilized in the various stage 3 (target interference) surveillance complexes. Class 1 systems generally utilize multiple Cas proteins to assemble their respective surveillance complexes. Among class 1

systems, the type-I and type-III systems are the most common, and can be differentiated by the presence of Cas3 and Cas10, respectively. Class 2 systems, on the other hand, generally utilize a single large protein subunit to assemble the surveillance complex, and among the class 2 systems, the type-II systems which utilize Cas9 are the most common. In addition, a number of "ancillary" proteins not directly implicated in the above processes are also clearly involved. Among these, proteins containing the CRISPR associated Rossmann Fold, or CARF domain [3] are central to this work.

In 2011 Lintner et al. studied two proteins with N-terminal CARF domains, Csa3 and Csx1, concluding these dimeric CARF domain proteins contained a single ligand binding site that spanned the dimer interface, and that the activities of these proteins would be regulated by a 2-fold (pseudo-) symmetric signal, potentially a cyclic oligonucleotide [4]. This hypothesis was subsequently echoed and expanded upon by others, as they identified HEPN nuclease domains in Csx1 and Csm6 similar to those in toxin-antitoxin systems, and suggested the unknown signal might activate these non-specific RNases. Non-specific degradation of cellular mRNAs by Csx1 and Csm6, in turn, are expected to deplete the cellular mRNA pool and inhibit protein synthesis. This, in turn, may result in the onset of cell dormancy until the infection can be cured, or even induce programmed cell death [3, 5].

In 2015 Yan et al. demonstrated manganese dependent 3'-deadenylase activity for yet another, apparently unrelated, ancillary Cas protein known as Csx3 [6]. They also determined crystal structures of Csx3 in the free form (3WZG), in the

presence of Mn^{2+} (3WZH), and in complex with a single stranded RNA fragment (3WZI). Interestingly, the 4-base RNA fragment was found distal from conserved active site histidines on the opposite face of this single domain protein. Topuzlu et al. then recognized Csx3 as a distant member of the CARF domain superfamily, and that the 4-base RNA fragment was bound in a nearly circular fashion in the putative regulatory binding site, suggesting a cyclic 4-base RNA as a regulatory ligand for CARF domain proteins [7].

In beautiful work, this hypothesis was then confirmed by both Kazlaukiene et al. [8] and Niewohner et al. [9]. They showed that when type-III surveillance complexes bind target RNA, the Cas10 cylcase domain synthesizes small oligoadenylates, including cyclic tetra- and hexa-adenylate (cA_4 and cA_6). Further, they also showed these products are recognized by the N-terminal CARF domains of Csx1 and Csm6, where they do indeed activate latent RNase activity within their C-terminal HEPN domains, consistent with the hypothesis that these cOA signals promote cell dormancy, or in the case that the cOA signal is not attenuated, programmed cell death.

Thus, what is the end fate of this cOA; are cellular cOA levels controlled reversibly? The answer to this conundrum was recently elucidated by Athukoralage et al., who found that some archaea express a “ring nuclease” that degrades the cA_4 signal into two molecules of A_2 [10]. Interestingly, this ring nuclease, christened “CRISPR ring nuclease 1” (Crn1), also utilizes a CARF domain for cA_4 recognition, but with catalytic residues installed to slowly degrade the cA_4 signal. Overall, these

advances in understanding the role of cyclic oligonucleotide signals in CRISPR-Cas prompted us to further investigate Csx3 and its potential interaction with the second messenger cA_4.

Results

Csx3 Associates with Type-III CRISPR-Cas Systems

In early seminal work on CRISPR-Cas, Haft et al. identified 45 different CRISPR associated (cas) gene families [11]. These included core cas genes common to multiple CRISPR types, as well as modular groups of cas genes now known to act in concert. However, clear contextual patterns for some cas families were elusive, and these cas genes were assigned to CRISPR subtype X, or csx. *Archaeoglobus fulgidus* (DSM4304) possesses both Type-IA and Type-IIIB CRISPR-Cas systems, with csx3 clearly embedded within the Type-IIIB cassette [12]. And while it has been noted that type-III systems may include ancillary csx3 genes [13], it is not clear if the reverse is true, i.e., that csx3 is most commonly associated with type-III systems, and if so, in what context. We thus identified putative AfCsx3 orthologs using a protein-protein BLAST search [14] against the non-redundant database. The top 46 unique sequences ranged from 84 to 113 residues in length and showed expect values $\leq 6 \times 10^{-17}$ with greater than 45% sequence identity to AfCsx3 over at least 75% of the query. Sequences below this threshold included larger, multi-domain proteins that, while containing a Csx3-like domain, are not clear homologs

and thus were not included. The sequence data was present in different forms, coming from 17 complete genomes, 7 scaffolds and 23 contigs. While the genomic context was not always clear, especially within contigs, csx3 was present in a type-IIIA cassette in 14 cases and in a type-IIIB cassette in 13 other cases. In just a single case was csx3 found outside a type-III system, where it was instead adjacent to type-I cassette. In this case, csx3 was at the extreme 5'-end of the contig, and we were unable to analyze the neighboring upstream sequence.

When the genes immediately up- or downstream of csx3 were surveyed, members of the csm6/csx1 family of CARF domain HEPN RNases were the most common, followed by cmr6, cmr4 and cas10 (Supplemental Figure 1). In contrast, except for a lone csa8a gene (type-IA), an immediate type-I or type-II neighbor was not observed. These single domain Csx3 orthologs thus show a clear association with type-III CRISPR-Cas systems. Because type-III (Cas10) surveillance complexes synthesize cOA signals when bound to target RNA, the association of Csx3 with type-III systems is consistent with the proposed recognition of cA_4.

Conserved Sequence Motifs in Csx3

We next looked for conserved sequences motifs in AfCsx3. A multiple sequence alignment performed with Clustal Omega [15] identified twelve strictly conserved residues (Supplemental Figure 1). These include two isolated residues, Pro^{33} and His^{80}, and three short motifs; GR(G/A)P(L/F)W^{50}, LxHxxH^{60} and DPR^{71}. When mapped to the AfCsx3 structure, His^{80} and the LxHxxH^{60} motif localize to the

RNase active site identified by Yan et al. [6]. In contrast, the GR(G/A)P(L/F)W^{50} and DPR^{71} motifs map to the tetra-base RNA binding site on the opposite face of the Csx3 dimer [6], and are structurally equivalent to the conserved motifs previously identified in the CARF domains of both Csa3 and Csx1 [4]. Thus, the putative cA_4 binding site is strictly conserved among this group of Csx3 orthologs.

Csx3 Binds cA_4 with Nanomolar Affinity

The presence of Trp^{50} among the conserved residues [GR(G/A)P(L/F)**W**^{50}] in the distal RNA binding site suggested intrinsic tryptophan fluorescence might be used to monitor ligand binding at this site. We utilized fluorescence lifetime spectroscopy for this purpose, taking advantage of an in-house time-resolved fluorescence spectrophotometer to collect subnanosecond-resolved fluorescence emission waveforms. This approach clearly demonstrates a very tight interaction between Csx3 and cA_4 (Figure 1). We measure an apparent K_d for cA_4 of 55 ± 4.5 nM. Notably, this value is less than the concentration of the Csx3 dimer (150 nM) used in the assay, explaining the near stoichiometric (linear) signal at cA_4 concentrations below 150 nM. Ideally, the assay would utilize Csx3 concentrations below the K_d, but the sensitivity of the assay does not currently allow this. The apparent K_d thus provides only an upper limit, and could be an underestimate of the true affinity. For perspective, a single cA_4 molecule per prokaryotic cell, approximate diameter of 1 μm, gives a concentration of ~ 100 nM. The affinity for cA_4 is thus highly relevant, allowing recognition even at concentrations

corresponding to a single molecule per cell. Clearly, Csx3 harbors a physiologically relevant cA_4 binding site.

We next asked how the affinity for cA_4 compares to the affinity for one of the poly-A tailed oligonucleotides originally utilized by Yan et al. For this, we chose the shortest of these, 5'-GGGAAAAGAAAAAA-3' [Supplemental Table 1, "A6" of Yan et al. [6]]. Here, the Csx3 fluorescence lifetime binding assay reveals a much lower affinity, giving a $K_d = 2.0 \pm 0.2$ μM (Figure 1). The significantly lower affinity for "A6" provides further evidence that cA_4 is indeed a physiologically relevant ligand.

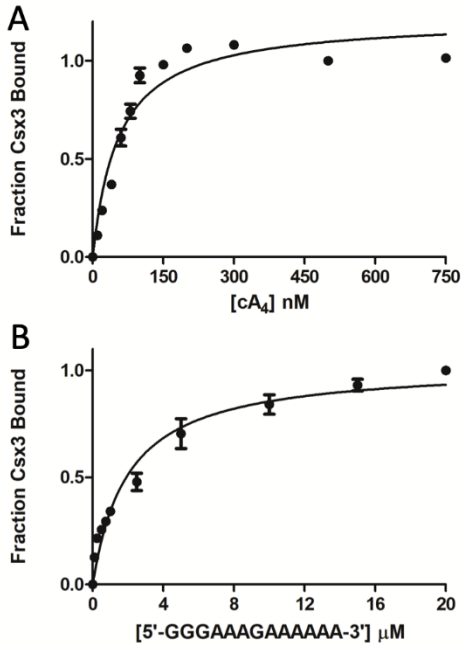


Figure 1. AfCsx3 binds cA4 with high affinity. A) The fluorescence lifetime of AfCsx3 Trp50 was used to follow cA4 binding. The mean waveform at each cA4 concentration was modeled as a linear combination of the cA4 bound and free waveforms, yielding the fraction bound. Each point is the mean of three independent experiments, each with three technical replicates. Error bars represent the standard deviation of the mean. The data were then fit to the one site single binding function in GraphPad Prism giving Kd = 55 ± 4.5 nM. Importantly, the

indicated cA4 concentrations along the abscissa represent total ligand concentration. Free ligand concentration is significantly less at the lower concentrations of cA4 used in this assay. Thus, the apparent Kd represents only an upper limit, and may well be an underestimate of the affinity. B) Note that the ligand concentrations in panel B are μM rather than nM as in panel A. The fluorescence lifetime binding assay shows significantly lower affinity for the small linear RNA with a poly-A tail, giving a Kd of 2.0 ± 0.2 μM.

The Csx3 Fold

To elucidate the structure of Csx3/cA_4 complex and conformational changes in in the Csx3 homodimer upon binding, we crystallized Csx3 in the presence of cA_4. Csx3 crystallized in space group I222 with a single subunit in the asymmetric unit, and usable diffraction data to 1.8 Å resolution. The structure unexpectedly revealed a 7th C-terminal ß-strand not present in previously modeled structures of AfCsx3 that runs anti-parallel to the first strand (Figure 2). Further, the additional ß-strand is preceded by a C-terminal α-helix (α3) that while present in the previously determined RNA bound structure (3WZI), is absent from the earlier apo- (3WZG) and Mn^{2+}-bound (3WZH) structures [6]. Thus, in contrast to previous crystal structures, the structural core of the Csx3 protomer is a mixed 7-stranded ß-sheet with ß7↑-ß1↓-ß2↑-α1-ß3↑-α2-ß4↑-ß5↓-ß6↑-α3 topology, in which the α-helices form right-handed crossovers between the indicated ß-strands (α3 connects ß6 to ß7). This places α1 and α2 on one face of the sheet, and α3 on the opposite side. In the context of the dimer, $α_1$ and α2 are buried at the subunit interface, while α3 is surface exposed.

It was unclear whether the presence of the additional ß-strand was due to the higher resolution of the data set, or whether the C-terminal tail is a dynamic structural element. To address this, we downloaded the 3WZI, 3WZG and 3WZH structure factors from the protein data bank, and computed molecular replacement structures using our new I222 structure as the search model, plus and minus the additional ß7-strand. In each case, the resulting R-factors were significantly improved (lower) for the search model lacking ß7. In addition, density for ß7 was absent in each of the respective maps. This suggests the C-terminal tail is mobile, and while present as a seventh beta-stand in this new I222 structure, adopts alternate conformations in the previous structures.

Table 1 - Data Collection and Model Refinement	
Data Set	AfCsx3
Wavelength (Å)	0.979
Space Group	I222
Cell Constants (a,b,c; Å) (α, β, γ; °)	53.45, 64.54, 70.11 90, 90, 90
Resolution Range[a] (Å)	46.25 - 1.80 (1.83 -1.80)
Unique Reflections[a]	10,996 (510)
Average Redundancy[a]	4.8 (4.1)
I/σ[a]	4.74 (2.87)
Completeness[a] (%)	99.5 (96.4)
R_{pim} (%)	3.2
R_{merge}[a] (%)	3.4 (4.4)
CC1/2[a]	0.999 (0.917)
Model	
R_{work}[a,b] (%)	18.7 (24.6)
R_{free}[a,b] (%)	21.6 (27.0)
Real Space CC[c] (%)	87.10
Mean B Value (overall; Å^2)	45.68
Coordinate Error (based on maximum likelihood, Å)	0.19
RMSD from ideality:	
Bonds (Å)	0.01
Angles (°)	0.82
Ramachandran Plot[d]:	
Most Favored (%)	98.04
Additional Allowed (%)	1.96
PDB Accession Code	6VJG
[a]Numbers in parenthesis refer to the highest resolution shell.	
[b]$R_{work} = \Sigma\|\|F_o\|-\|F_c\|\|/\Sigma\|F_o\|$ where F_o and F_c are the observed and calculated structure factor amplitudes used in refinement. R_{free} is calculated as R_{work}, but using the "test" set of structure factor amplitudes that were withheld from refinement (9.97%).	
[c]Correlation coefficient (CC) is agreement between the model and $2mF_o$-DF_c electron density map.	

Table 1. Data Collection and Model Refinement

Further examination of this I222 structure revealed that ß7 is anchored not only by hydrogen bond interactions as it runs antiparallel to ß1 of the same subunit, but that it is also stabilized by a parallel ß-strand interaction with ß6 of a neighboring subunit (Figure 3). Interestingly, this results in formation of an extended, inter-subunit ß-sheet running the length of the crystal. Visual observation and PISA interface analysis software [16], however, indicate the "open symmetry" of this inter-subunit interaction is insufficient to support formation of a stable protein interface in solution. While such open symmetry is common in filamentous protein assemblies, it's atypical for enzyme complexes, which more commonly show closed point group symmetry. Regardless, the dynamic nature of the C-terminal tail and its ability to mediate inter-dimer interactions is reminiscent of 3D domain swapping [17] and may indicate the ability of the C-terminal tail to mediate additional protein-protein interactions within the cell.

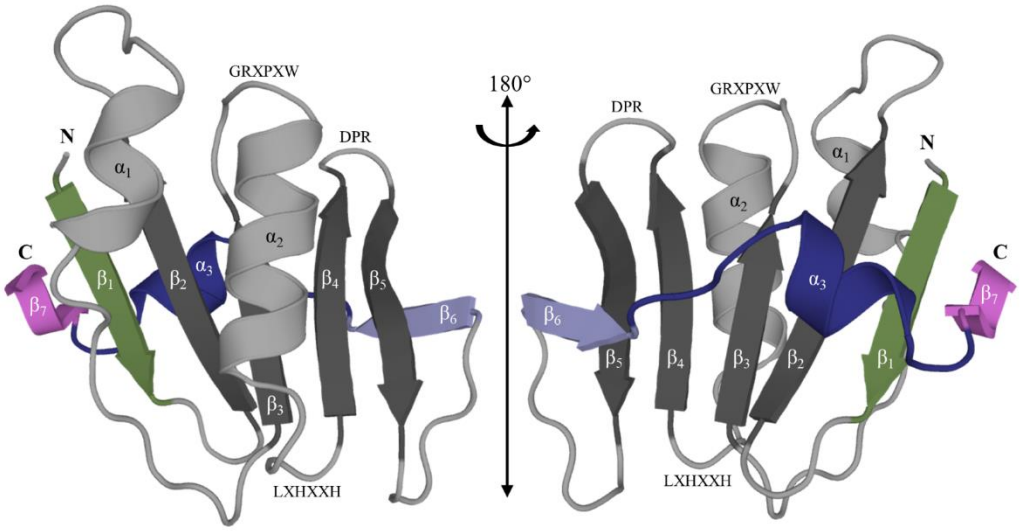


Figure 2. The AfCsx3 Fold. Previous structures of Csx3 reported a 6-stranded mixed ß-sheet with 2 α-helices. The high-resolution crystal structure described here revealed an ordered C-terminus with a third alpha helix (α3, dark blue) followed by an additional ß-strand (ß7, violet). This C-terminal extension crosses the outside, solvent exposed surface of the ß-sheet, allowing ß7 to add antiparallel to ß1. The fold is thus a mixed seven-stranded ß-sheet with ß7↑-ß1↓-ß2↑-α1-ß3↑-α2-ß4↑-ß5↓-ß6↑-α3 topology with the α-helices providing right-handed crossovers between the indicated ß-strands. In a dimer, α1 and α2 are buried at the subunit interface while α3 is exposed on the surface. Conserved motifs GRXPXW, LXHXXH and DPR are found in the ß3-α2, α2-ß4, and ß4-ß5 loops. Note the LXHXXH RNase active site motif that lies at the "bottom" of the protomer, while the GRXPXW and DPR motifs involved in cA_4 recognition are at the top.

Although Csx3 was co-crystallized with cA_4, density for cA_4 was not observed in the RNA binding pocket, or elsewhere. While Csx3 is a Mn^{2+} dependent deadenylase, our crystallization conditions did not contain added divalent cations. Thus, given the extreme affinity of Csx3 for cA_4 and the previous structure of Yan et al. with an unidentified RNA fragment (3WZI), this was unexpected. On the other hand, crystallization conditions did not include chelators and trace metals could have enabled cA_4 degradation by the RNase active site. Alternatively, self-

regulating metal independent ring nuclease activity has been seen for select CARF domains [18-20], suggesting Csx3 might also harbor metal independent, cA_4 specific ring nuclease activity at the cA_4 binding site, as opposed to the metal dependent RNase active site present on the opposite face of the protein.

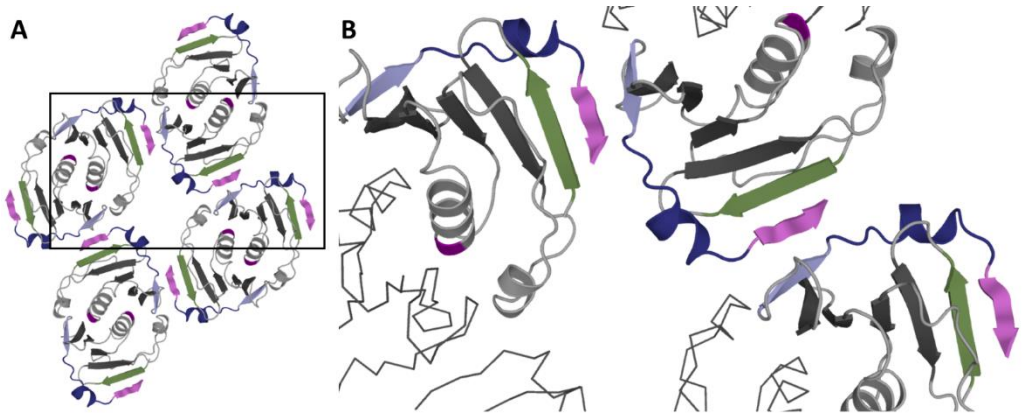


Figure 3. The dynamic C-terminal tail. A) The Csx3 dimer packs in sheets or planes within the crystal. B) The inter-dimer crystal contacts within a plane are anchored primarily by a parallel ß-strand hydrogen bond interaction between $ß_7$ (violet) of one subunit and $ß_6$ (light blue) of the neighboring dimer. The ordered ß7-strand of the C-terminal tail thus serves to cross link dimers within the crystal as it "inserts" between ß1 (green) of one subunit and ß6 of the neighbor. This interaction leads to the formation of continuous extended ß-sheets that running the length of the crystal. Visual observations and PISA interface analysis software indicate these interactions are not substantial enough to support the formation of a stable protein interface in solution. This open symmetry is thus a likely artifact of crystallization conditions and is not indicative of interactions *in vivo*. That said, the dynamic C-terminal tail clearly has the ability to form inter-protein interactions which might mediate protein-protein interactions within the cell.

Ring Nuclease Activity

For these reasons, we examined the ability of Csx3 to degrade cA_4. The activity assay utilized thin layer chromatography to separate unlabeled cA_4 (Biolog Life Science/Axxora) from potential degradation products. Using this assay, we

observed unexpectedly high levels of ring nuclease activity upon the addition of 200 μM Mn^{2+} (Figure 4), where the thin layer chromatography was consistent with diadenylate (A_2) as the major product [10]. At times trace activity levels were observed in samples lacking added Mn^{2+}, perhaps due to low-level copurification of Csx3 with divalent cations, and potentially explaining the lack of bound cA_4 in the crystal structure. In general, this activity was inhibited upon addition of EDTA. Importantly, when the progress curve in the presence of 200 μM Mn^{2+} was modeled as first order exponential decay, the average observed rate constant (k_{obs}) from 5 independent experiments was 4.95 ± 0.43 min^{-1}. For perspective, the rate constant for CRISPR ring nuclease Crn1 identified by Athukoralage et al. in 2018 is 0.089 min^{-1} [21, 22], more than an order of magnitude slower than Csx3 in the presence of Mn^{2+}. From this perspective Csx3 demonstrates robust cyclic oligonucleotide phosphodiesterase, or "ring nuclease" activity. This suggests cA_4 is indeed a physiologically relevant substrate and that Csx3 functions within CRISPR-Cas to regulate the amplitude and duration of the cA4 signal.

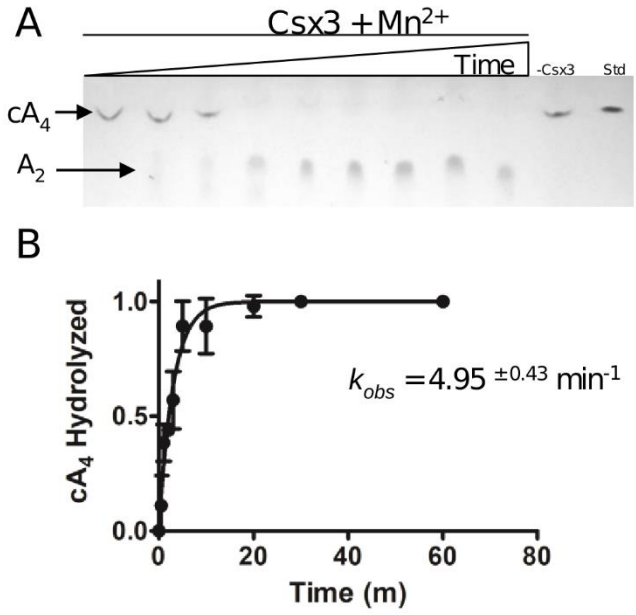


Figure 4. Af Csx3 ring nuclease activity. A) TLC analysis showing a representative time course for the degradation of unlabeled cA_4 (150 µM) incubated with Af Csx3 (10 µM dimer) at 50 °C in the presence of 200 µM Mn^{2+}. Lanes 1-9 are time points at 0, 0.5, 1, 3, 5 10, 20, 30, and 60 minutes. Lane 10 was a 60 minute incubation in the presence of Mn^{2+} in the absence of Csx3. Lane 11 was loaded with 1.5 nanomoles of cA_4 standard. B) Data from five independent cA_4 ring nuclease activity replicates were averaged, and the data fit to a first order exponential decay model yielding an apparent first order rate constant, k_{obs} = 4.95 per minute. The standard deviations at each time point are represented by error bars.

Discussion

Defining Elements of the Cyclic oligoAdenylate Recognition Fold (CARF)

With confirmation that Csx3 does indeed bind cyclic tetra-adenylate with high affinity, we can now compare and contrast the structure of this more distant member of the CARF domain superfamily with the canonical CARF domain fold. This is well represented by the relatively simple *S. solfataricus* Csa3 CARF domain [4], which adopts the classic dinucleotide binding domain fold with one simple

modification; the right handed helical crossover between ß5 and ß6 is lost. The two ß-strands are instead connected by a reverse turn, with ß6 running antiparallel rather than parallel to ß5. This modification has significant consequences. As Csa3 forms the homodimer, the helices connecting ß4 and ß5 in each subunit pack against and parallel to each other at the center of the dimer interface, with each assuming the place of the "missing" helix in the neighboring subunit. The ß5-ß6 reverse turn and loss of the connecting helix is thus responsible for creation of the CARF domain dimer interface (Figure 5). And because dimer formation is inherent to recognition of the 2-fold symmetric cOA signals, this is a defining characteristic of the CARF domain fold. In short, the ß4↑-α4-ß5↑-ß6↓ super-secondary structural element in the Csa3 CARF domain lies at the heart of the fold. Critically, these structural elements in Csa3 (Csm6 and Csx1 as well) are topologically analogous to the ß3↑-α2-ß4↑-ß5↓ elements in Csx3, and within the Csx3 dimer they play the same central role at the dimer interface as the cognate elements in Csa3 and other CARF domain proteins.

As the CARF domain acronym (CRISPR Associated Rossmann Fold) went into use before the ligands for this domain were fully identified, it does not well convey the function of this fold. Now that it is understood to bind small cyclic oligonucleotides, it is tempting to give it a more meaningful name. However, the CARF acronym is now in such wide use that renaming it could be difficult. We note however, that this domain clearly serves as a Cyclic oligo-Adenylate Recognition Fold, and this mnemonic could help some to remember its function. One should keep in mind, however, that it also recognizes small linear polynucleotides [6], and

it is difficult to imagine that nature has not also adapted this fold for the recognition of other bases.

The C-terminal Tail

The previous structures of Csx3 by Yan et al. show a disordered C-terminal tail. In the structure reported here, the C-terminal tail is instead captured in an ordered conformation. The structure shows a well-defined C-terminal α-helix that traces a path along the "back side" of the subunit to position the extreme C-terminus to add an additional ß-strand along the outside edge of the first ß-strand (Figure 2). Thus, we see that the C-terminal tail not only has a propensity to form these secondary structural elements, but to utilize them, especially ß7, to interact with other secondary structural elements within Csx3. However, the dynamic nature of the tail suggests that within the cell, the C-terminal tail might instead be utilized to interact with other cellular components, potentially as a member of a larger complex that might modulate the activity of Csx3. Importantly, in this conformation both ß1 and ß7 are available for intersubunit interactions.

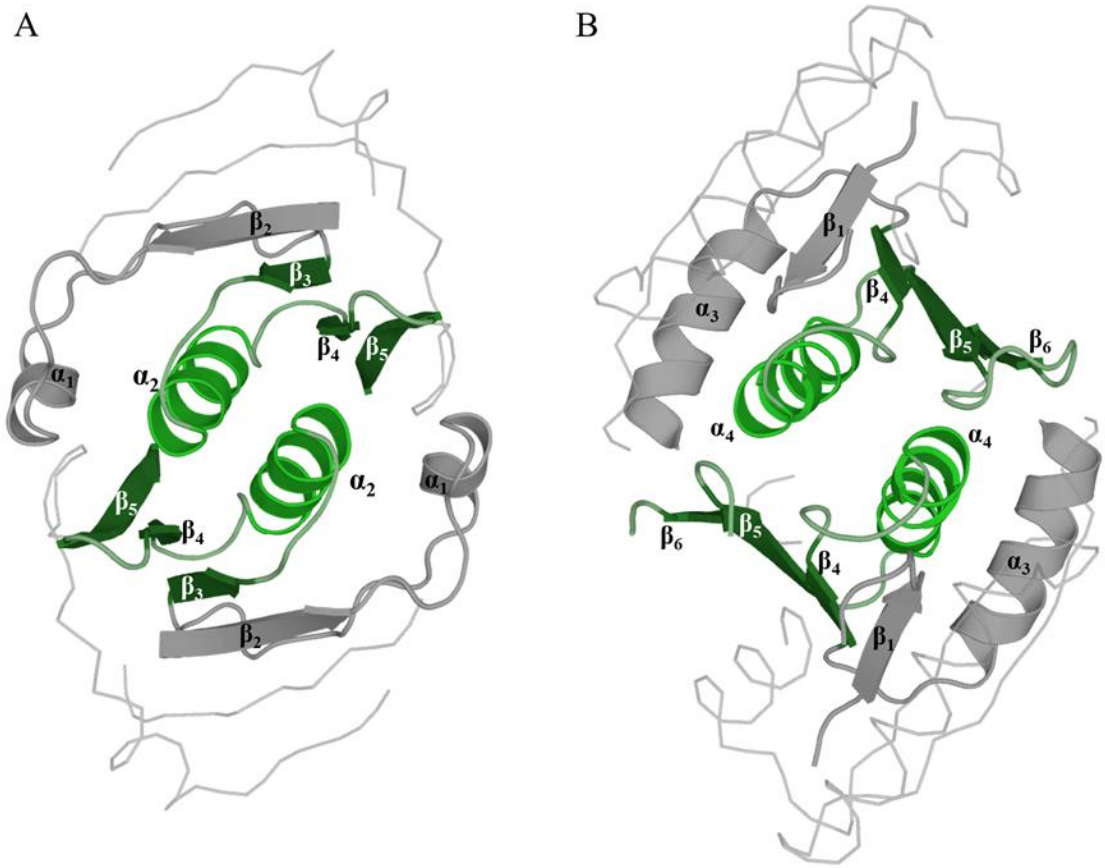


Figure 5. Structural definition of the CARF domain core. Superposition of the ß3-α2-ß4-ß5 sequence of Csx3 (A) onto the ß4-α4-ß5-ß6 sequence of *S. solfataricus* Csa3 CARF domain (B) (RMSD 2.9 Å) reveals core secondary structures shared by the two proteins (green). Csa3 and Csx3 each contain this contiguous ßαßß super secondary structures in which the first two ß-strands are connected by a right-handed helical crossover followed by a reverse turn into the 3rd ß-strand. The core helix in this motif is an essential part of the dimer interface, it packs against the two beta-strands connected by the reverse turn, $ß_4ß_5$ in Csx3 (left) and $ß_5ß_6$ in Csa3 (right). In the classic dinucleotide binding domain, these beta-strands are instead connected by a right-handed helical crossover, in which the connecting helix occupies the same space as the helix from the adjacent Csa3 or Csx3 subunit. The reverse turn connecting these ß-strands is thus a critical structural feature of the CARF domain dimer interface. In addition, Csx3 places ß2 and α1 in positions equivalent to the discontinuous ß1 and α3 secondary structural elements of Csa3 (grey).

Recognition of cA_4

The core ß3↑-α2-ß4↑-ß5↓ structural motif is functionally important as well. The conserved GR(G/A)P(L/F)W^{50} and DPR^{71} motifs identified above lie in the ß3α2 and ß4ß5 loops, respectively, and play a significant role in construction of the Csx3 RNA binding pocket (Figure 5). Similarly, conserved residues are also found at these positions in Csa3 and Csx1, where they are also involved in construction of the conserved ligand binding site [4].

Additional insight into recognition of cA_4 by Csx3 can be gained by re-examining the RNA density in the 3WGI structure determined by Yan et al [6]. The 3WGI crystals were grown in the presence of a 37 nucleotide single-stranded RNA molecule, but Yan et al. report that the RNA was degraded over the time course for crystal growth. The 4 ordered nucleotides in the CARF domain binding site were thus modeled as a linear 4-base RNA fragment.

However, perhaps because the density arises from the presence of RNA fragments bound in overlapping conformations with different register, the electron density map clearly shows a cyclic ring of electron density. This suggested the density might be used to inform efforts to model the Csx3/cA_4 interaction. Indeed, when we build cA_4 into this density, our minimally refined cA_4 bound model gives R_{work} and R_{free} values of 19.2 and 22.4% respectively, which compares very favorably to the original model with the linear 4-base RNA fragment (20.7/23.5%). Thus, while the 3WGI structure does not stem from crystals of Csx3 in complex with cA_4, the ring like density is well modeled by cA_4 (Figure 6).

As expected, this model shows the conserved GR(G/A)P(L/F)W[50] and DPR[71] motifs in the ß3α2 and ß4ß5 loops play a central role in cA_4 recognition. In addition to these elements, the ß2α1 loop is also prominent, corresponding to the conserved ß1α1 loop of Csa3. Analysis of the modeled interaction shows that cA_4 buries greater than 800 $Å^2$ of protein surface. Base A1 (and A3 in the second subunit) is nestled in a hydrophobic pocket between Arg[46], Tyr[68] and Pro[70], with potential hydrogen bonds from the main chain NH of Gly[45], the His[15] side chain and Arg[46] to N1, N6 and N7 of the adenine ring, respectively. Base A2 (and A4) stacks on top of Pro[48] and the edge of Trp[50], with N6 donating a hydrogen bond to the main chain carbonyl of Ile[22]. The phosphate groups in this model lack substantial interactions, though Arg[46] and Arg[71] could seemingly interact strongly with a relatively minor reorientation of their side chains. Notably, the phosphate groups adjacent to the Arg[46] side chains are elevated above the plane of the cA_4 ring while those adjacent to Arg[71] are more deeply buried. Finally, strictly conserved Asp[69] hydrogen bonds to the 2'-OH of the ribose of A2 (and A4).

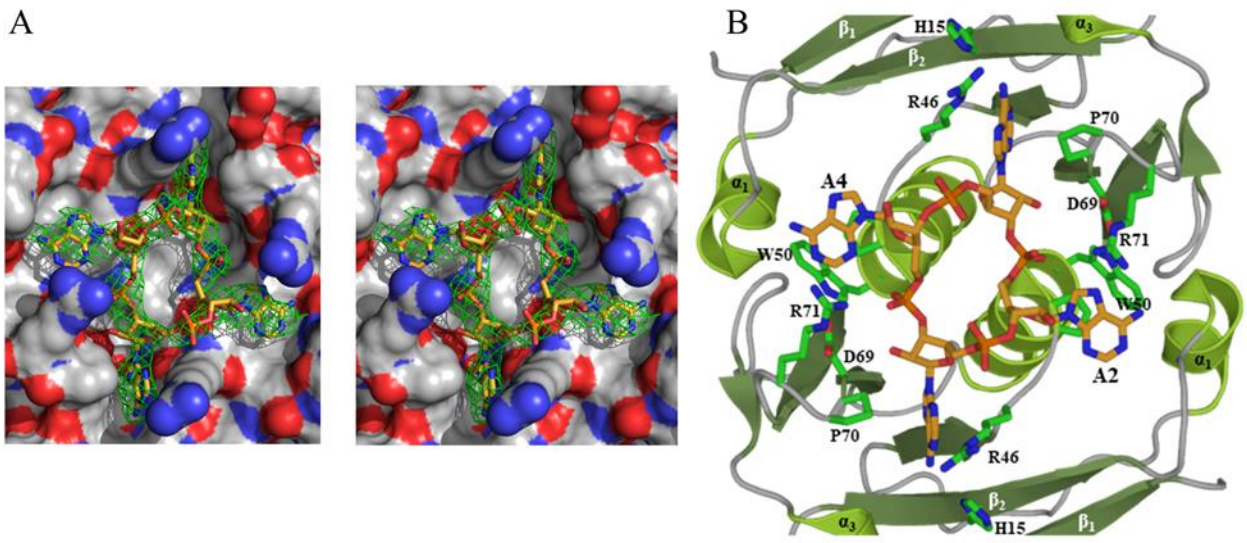


Figure 6. RNA density in the 3WZI Csx3 structure accommodates cA4. A) Stereo image. An omit map was generated using the 3WZI model with RNA atoms deleted and contoured at 1.5 σ (green mesh). The omit map was then used to inform modeling the cA_4/Csx3 interaction. Csx3, oriented as in panel 5A, is represented as a surface model with nitrogen atoms in blue, oxygen atoms in red, carbon atoms colored grey, and cA_4 is represented in sticks with carbon atoms colored yellow. B) Csx3 secondary structure elements shown in cartoon representation with select side chains shown as sticks. cA_4 is also shown in stick representation with carbons in beige.

Structural Basis for Ring Nuclease Activity

Yan et al. quite logically proposed a tethered substrate model for Csx3 in which the RNA was bound in the RNA binding site on one face of the Csx3 dimer, and then extended into the active site on the opposite side of the subunit for cleavage. However, this model cannot explain the degradation of cA_4, which is far too small. The CRISPR associated HEPN RNases such as Csx1 and Csm6 are two domain proteins that utilize an N-terminal CARF domain to bind cyclic oligoadenylates that then activates the C-terminal HEPN RNase domains. In some cases, these CARF domains have catalytic residues installed in the cA_4 binding site

that slowly degrade the cA_4 signal in a metal independent mechanism [18-20]. In this sense, the activities of these ancillary Cas proteins are "self-limiting", much like the classic picture for slow GTP hydrolysis by G-proteins.

We might consider a similar model for Csx3. The crystal structure suggests Asp^{69} in the cA_4 binding site could potentially act as a general base to deprotonate water or the ribose 2'-OH, catalyzing nucleophilic attack on the neighboring phosphorus to cleave cA_4, resulting in a product with 3' phosphate groups. However, this mechanism fails to explain the effect of Mn^{2+} on the reaction, which instead points to the histidine and metal rich RNase active site on the other face of the molecule [6]. One alternative solution is to invoke a metal independent RNase active site, though this is inconsistent with the data of Yan et al. [6] and our own ring nuclease data. Another possibility is a mechanism that utilizes both sites. Whether this might involve activation of the metal dependent active site [6] by binding of cA_4 at the allosteric regulatory binding site, or some other mechanism is not clear. The structure of the RNase active site seems largely unchanged by the presence of the 4-base RNA fragment, but this fragment was a degradation product, which would not be expected to activate the CARF domain. In any event, we currently lack a satisfying explanation for this activity and additional work is clearly needed.

Utility of Unlabeled cA_4 Ring Nuclease and Fluorescence Lifetime Binding Assays

Previous work on CARF domain proteins has employed radiolabeled substrates and ligands to measure both ring nuclease activity and binding affinity for cA_4. Our work, in contrast, has utilized unlabeled cA_4 for both the ring nuclease and fluorescence lifetime binding assays. The "cold" ring nuclease assay reported here is easily adapted for the study of ring nuclease activity in general, and thus of general use to the field. And while the fluorescence lifetime binding assay is specific to proteins containing a tryptophan residue in the ligand binding site, Trp^{50} is strictly conserved among the Csx3 orthologs identified in this study and is thus applicable to Csx3 in general. Notably, the assay benefits from the stoichiometry of the interaction. With a single cA_4 molecule bound across the dimer interface, the change in the fluorescence waveform arises from two copies of Trp^{50} for each cA_4 bound. This label free binding assay could be utilized for the development of CARF domain agonists and antagonists, which may have potential therapeutic applications in pathogenic bacteria carrying type III CRISPR-Cas systems or other CARF domain proteins.

Physiological Specificity for cA_4

In a significant advance, Csx3 was previously described as a manganese dependent deadenylation ribonuclease, with specificity for poly-A tails in single stranded RNAs [6]. While a rate constant for this activity was not reported, we note the assay conditions utilized a 2:1 ratio of Csx3 (20 μM) to substrate RNAs (10 μM),

and that under these conditions, significant uncleaved substrate is still seen at the 5-minute time point. In contrast, we have instead used cA_4 in excess, with a Csx3:cA_4 stoichiometry of 1:15 (10 µM Csx3/150 µM cA_4), and under these conditions, cA_4 is largely degraded within 5 minutes. This simple comparison suggests Csx3 shows at least 20 times greater specific activity for cA_4 than it does for small poly-A tailed RNAs. Further, the dissociation constant for the poly-A tailed RNA reported here indicates Csx3 shows ~50 fold greater affinity for cA_4 than this small poly-A RNA. Considering both the 50-fold lower K_d and the 20-fold greater catalytic activity, the overall catalytic efficiency of Csx3 is at least 1,000-fold greater for cA_4 than it is for the poly-A tailed RNA substrate. These are strong arguments that cA_4 is indeed the physiologically relevant substrate, and that Csx3 functions within the cell as a cyclic oligonucleotide phosphodiesterase, or ring nuclease.

Additional evidence for a physiological role for Csx3 as a ring nuclease comes from consideration of the activities seen for other CRISPR associated ring nucleases [10]. Specifically, Csx3 ring nuclease activity is an order of magnitude greater than that seen for Crn1, and is approaching the level of activity seen for virally encoded anti-CRISPR (Acr) ring nucleases [21]. If anything, the activity is perhaps higher than expected [23], and its activity may need to be regulated at some level if present in the cell at significant concentrations. In this regard, the ability of the dynamic C-terminal tail to mediate interactions within a larger complex may be extremely relevant.

Overall then, Csx3 possesses the ability to recognize cA_4 at levels as low as a single molecule per cell and is also well endowed with ring nuclease activity. Once an infection is cured, and Cas10 ceases production of cA_4, these attributes would allow it to quickly and effectively scrub the cell of cA_4. Csx3 thus has all the attributes needed to serve as an effective counterbalance to cA_4 production by Cas10, allowing it to regulate the duration and amplitude of the cA_4 signal. In cases where the infection is cured, the cell can then return to normal metabolic activity, providing an off-ramp from the programmed cell death pathway. Importantly, ring nucleases have yet to be identified in many organisms harboring type III systems. The assignment of this activity to Csx3 thus represents a significant advance in our understanding of CRISPR-Cas in cells utilizing type III CRISPR-Cas systems.

Experimental Procedures

Gene Annotations

Protein accessions were matched to their respective genome accessions. In the case of "non-redundant" Csx3 sequences, a representative genome with the highest level of completion was analyzed. In cases where a contig or scaffold were the only available information, analysis was restricted to the assembled read. Genome accessions were downloaded from the GenBank database with annotations from NCBI. In each case, *csx3* was already annotated. Genomes were then opened and viewed by eye using the Unipro UGENE software [24]. For genomes in which the neighboring genes were already annotated, those annotations were used. In the

event that neighboring genes lacked annotation, the amino-acid annotation was analyzed with HHpred [25]. Identification was considered conclusive only when the E-value was <0.001 and probability was greater than 95%.

Csx3 Expression and Purification

The native coding sequence for AfCsx3 with a minimal N-terminal His_6-tag (Met-His_6) was purchased from Thermo Fisher Scientific and moved into Gateway plasmid pDEST14 using previously described protocols [26] to give the pDEST14-AfCsx3 expression vector. For protein expression, *E. coli* BL21 (DE3)-pRIL (Stratagene) was transformed with pDEST14-AfCsx3. Typically a single colony was used to inoculate a 5 ml overnight LB starter culture (100 µg/ml ampicillin and 34 µg/ml chloramphenicol), which was then used to inoculate ZYP-5052 autoinduction media [27] (1:1000) and grown at 20°C for 48 hours. Cells were harvested by centrifugation at 4,000 x g for 30 min and 5-7 g pellets were stored at -20 ^{o}C until needed. Cell pellets were then thawed and resuspended in 5 mL lysis buffer (20 mM Tris-Cl pH 7.5; 500 mM NaCl; 5% glycerol; 1 mM PMSF) per gram of cell pellet and lysed by French Press. The lysate was centrifuged at 25,000 x g for 30 minutes and the supernatant applied to a Ni-NTA affinity column, which was then washed with lysis buffer plus 20 mM imidazole and eluted with lysis buffer plus 500 mM imidazole. Fractions containing Csx3 were then combined and applied to a Superdex S-75 column equilibrated with 10 mM Tris-Cl pH 7.5, 500 mM NaCl, 5% glycerol. Protein concentrations were determined by Bradford assay [28] using

Protein Assay Reagent (Bio-Rad) and BSA as a standard. The purity and molecular weight of Csx3 were confirmed by SDS-PAGE. Standard yields were approximately 25 mg of pure AfCsx3 from a 5 g cell pellet. Fractions containing pure Csx3 were combined and aliquots at 1 mg/ml were stored at -20 °C.

Fluorescence Lifetime Binding Assays

cA_4 was titrated (0 – 750 nM) across a 96-well quartz microplate (Hellma) containing 150 nM Csx3 dimer in 20 mM Tris-Cl pH 7.5, 100 mM NaCl, and 1.0% glycerol (final concentrations) with total well volumes of 350 µl, with three technical replicates per plate for each experiment. Binding assays were also performed with Poly-A Tail RNA (5'-GGGAAAAGAAAAAA-3') as described above using 300 nM Csx3 dimer and 0 – 20 µM RNA. Wells were excited at λ = 295 nm and time-resolved emission of individual wells followed at λ = 350 nm using a NovaFluor fluorescence lifetime spectrometer [Fluorescence Innovations, [29, 30]]. Waveforms were recorded over a total decay time of 128 ns at 2 ps intervals. The fluorescence lifetime data were analyzed as described in supporting information of Schlick et al. [31]. Briefly, the waveforms collected from each well were fit to a linear combination of the free (W_f) and complexed (W_c) waveforms, yielding the fraction of protein in complex with ligand at each specific cA_4 concentration. The data were then fit to the "one-site specific binding" function in GraphPad Prism ($Y = Bmax*X/(K_d + X)$).

Crystallization and Structure Determination

For crystallization, AfCsx3 was thawed and exchanged into crystallization buffer (20 mM Tris-Cl pH 7.5; 125 mM NaCl; 5% glycerol; 500 mM imidazole) using an S-75 Superdex size exclusion column, and then concentrated to 7 mg/mL using a centrifugal spin concentrator (Amicon Millipore 10K). Crystals of AfCsx3 were grown at 23°C using hanging drop vapor diffusion with a well solution containing 100 mM Tris-Cl pH 8.5, 200 mM Ammonium Phosphate, 45% (+/-)-2-methyl,2,4-pentanediol. Drops were assembled with 1 µl of protein, 1 µl of well solution and 0.1 µl of 10 mM cA_4 (c-tetraAMP, BioLog Lifescience Institute). Crystals grew to a maximum of 150 x 150 x 100 µm over the course of 3-4 weeks. The crystals were looped and flash frozen in liquid N_2. X-ray diffraction data were collected at APS-Argonne National Laboratory NE-CAT 24-ID-C. Data were indexed, integrated and scaled in space group I222 using the HKL2000 software package [32]. The Phaser-MR module [33] of Phenix [34] was used to determine initial phases for Csx3 using the existing structure of apo-Csx3 (PDB ID 3WZG) determined by Yan et al. [6]. The structure was completed by iterative model building with Coot [35] and refinement with phenix.refine [36]. The final R_{work} and R_{free} were 18.7% and 21.6%, respectively. MolProbity was used for model validation [37], with 98% of residues falling in the most favored regions of the Ramachandran plot and none in disallowed regions. Additional statistics on data and model quality, and model refinement are presented in Supplemental Table 1. Both the model and structure

factors have been deposited in the Protein Data Bank (6VJG). Structural Figures were generated with PyMol [38].

Cyclic Oligonucleotide Phosphodiester-ase Assays

"Ring nuclease" assays were adapted from those of Athukoralage et al., but instead utilized unlabeled cA_4 [10]. Briefly, cA_4 (c-tetraAMP, BioLog Lifescience Institute, final concentration 150 μM) was added to reactions containing 10 μM AfCsx3 in 10 mM Tris-Cl pH 7.5, 200 μM $MnCl_2$, 500 mM NaCl and 5% glycerol (final concentrations) plus SUPERase•In inhibitor and incubated at 50 ^oC. 10 μl aliquots were removed from the reaction mixture and quenched with equal volume phenol-chloroform at the desired time points. Upon quenching, 20 μl H_2O was added to each aliquot to bring the sample volume up to 30 μl aqueous and 10 μl organic. 20 μl of the aqueous portion were then removed and placed in a SpeedVac for 30 minutes or until entirely evaporated. Samples were then re-suspended in 20 μl methanol, centrifuged in a microcentrifuge at 14,000 RPM for 10 minutes to concentrate any precipitate, and the top 15 μl was removed and again placed in a SpeedVac for 30 minutes or until evaporated, re-suspended in 5 μl methanol, and spotted onto a glass-backed F254 silica gel TLC plate. The TLC plate was developed in 70% ethanol and 0.2 M ammonium bicarbonate (pH 9.3) until the mobile phase reached approximately 75% of the plate height. Plates were dried and imaged with an F254 lamp and imaging box. The intensities of the residual cA_4 bands were quantitated using ImageJ software and normalized to time = 0. Using GraphPad

Prism, the data were then fit to progress curves as described by Sternberg et al. [39] to determine the first order rate constants (k_{obs}).

Data Availability

The Csx3 model and associated X-ray diffraction data (structure factors) have been deposited in the Protein Data Bank (6VJG).

Funding and Additional Information

Research reported in this publication was supported by the National Science Foundation grant MCB-1413534 to CML. SB acknowledges support by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103474. Fluorescence assays were collected at Montana State University on instruments funded by the National Science Foundation grant MRI-0821802. X-ray diffraction data were collected in the Macromolecular Diffraction Laboratory at Montana State University supported by the MJ Murdock Charitable Trust and the Northeastern Collaborative Access Team beamlines, which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Eiger 16M detector on 24-ID-E is funded by a NIH-ORIP HEI grant (S10OD021527). This research also used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH1

References

1. Sorek, R., C.M. Lawrence, and B. Wiedenheft, *CRISPR-mediated adaptive immune systems in bacteria and archaea.* Annu Rev Biochem, 2013. **82**: p. 237-66.

2. Hille, F., et al., *The Biology of CRISPR-Cas: Backward and Forward.* Cell, 2018. **172**(6): p. 1239-1259.

3. Makarova, K.S., et al., *CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems.* Front Genet, 2014. **5**: p. 102.

4. Lintner, N.G., et al., *The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System.* J Mol Biol, 2011. **405**(4): p. 939-955.

5. Makarova, K.S., et al., *Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes.* Biol Direct, 2012. **7**: p. 40.

6. Yan, X., W. Guo, and Y.A. Yuan, *Crystal structures of CRISPR-associated Csx3 reveal a manganese-dependent deadenylation exoribonuclease.* RNA Biol, 2015. **12**(7): p. 749-60.

7. Topuzlu, E. and C.M. Lawrence, *Recognition of a pseudo-symmetric RNA tetranucleotide by Csx3, a new member of the CRISPR associated Rossmann fold superfamily.* RNA Biol, 2016. **13**(2): p. 254-7.

8. Kazlauskiene, M., et al., *A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems.* Science, 2017. **357**(6351): p. 605-609.

9. Niewoehner, O., et al., *Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers.* Nature, 2017. **548**(7669): p. 543-548.

10. Athukoralage, J.S., et al., *Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate.* Nature, 2018. **562**(7726): p. 277-280.

11. Haft, D., et al., *A Guild of 45 CRISPR-Associated (Cas) Protien Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.* Plos Computational Biology, 2005. **1**(6): p. 474-483.

12. 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.

13. Shah, S.A., et al., *Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families.* RNA Biol, 2019. **16**(4): p. 530-542.

14. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

15. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W636-W641.

16. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.

17. Bennett, M.J., M.P. Schlunegger, and D. Eisenberg, *3D domain swapping: a mechanism for oligomer assembly.* Protein Sci, 1995. **4**(12): p. 2455-68.

18. Athukoralage, J.S., et al., *A Type III CRISPR Ancillary Ribonuclease Degrades Its Cyclic Oligoadenylate Activator.* J Mol Biol, 2019. **431**(15): p. 2894-2899.

19. Garcia-Doval, C., et al., *Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6.* Nat Commun, 2020. **11**(1): p. 1596.

20. Jia, N., et al., *CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity.* Mol Cell, 2019. **75**(5): p. 944-956 e6.

21. Athukoralage, J.S., et al., *An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity.* Nature, 2020. **577**(7791): p. 572-575.

22. Rouillon, C., et al., *Control of cyclic oligoadenylate synthesis in a type III CRISPR system.* Elife, 2018. **7**.

23. Athukoralage, J.S., et al., *The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling.* Elife, 2020. **9**.

24. Okonechnikov, K., et al., *Unipro UGENE: a unified bioinformatics toolkit.* Bioinformatics, 2012. **28**(8): p. 1166-7.

25. Zimmermann, L., et al., *A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.* J Mol Biol, 2018. **430**(15): p. 2237-2243.

26. Menon, S.K., et al., *The crystal structure of D212 from sulfolobus spindle-shaped virus ragged hills reveals a new member of the PD-(D/E)XK nuclease superfamily.* J Virol, 2010. **84**(12): p. 5890-7.

27. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein expression and purification, 2005. **41**(1): p. 207.

28. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Anal Biochem., 1976. **72**: p. 248-54.

29. Petersen, K.J., et al., *Fluorescence lifetime plate reader: resolution and precision meet high-throughput.* Rev Sci Instrum, 2014. **85**(11): p. 113101.

30. Muretta, J.M., et al., *High-performance time-resolved fluorescence by direct waveform recording.* Rev Sci Instrum, 2010. **81**(10): p. 103101.

31. Schlick, K.H., et al., *Characterization of protein aggregation via intrinsic fluorescence lifetime.* J Am Chem Soc, 2009. **131**(46): p. 16608-9.

32. Otwinowski, Z. and W. Minor, *[20] Processing of X-ray diffraction data collected in oscillation mode*, in *Methods in Enzymology*. 1997, Academic Press. p. 307-326.

33. McCoy, A.J., et al., *Phaser crystallographic software.* Journal of Applied Crystallography, 2007. **40**: p. 658-674.

34. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallogr D Biol Crystallogr, 2010. **D66**: p. 213-221.

35. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 486-501.

36. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D-Structural Biology, 2010. **66**: p. 213-221.

37. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallographica Section D-Structural Biology, 2010. **66**: p. 12-21.

38. DeLano, W.L. *The PyMOL Molecular Graphics System*. *http://www.pymol.org*. 2002; Available from: http://www.pymol.org.

39. Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna, *Mechanism of substrate selection by a highly specific CRISPR endoribonuclease*. RNA, 2012. **18**(4): p. 661-72.

CHAPTER SIX

CONCLUSION

The research described in this dissertation spans two very different areas of research linked by methodology used to overcome the bottleneck or rate limiting step of generating significant quantities of protein for characterization. The research presented in Chapter Three and Appendix C describes investigations of regulatory mechanisms in the Class 1 CRISPR-Cas immune response that is yet to be fully understood. Chapter Four describes investigations leading to the development of a highly efficient method for the expression and purification of Galectins, and explains its use in a CURE-based environment.

Conclusions from Chapter Three

The functional and structural characterization of the *S. solfataricus* Csa3a/cA_4 complex found in Chapter Three identifies a nexus linking the immune responses of the distinctly separate Type I and Type III CRISPR-Cas systems [1-5]. While earlier studies proved CARF superfamily proteins specifically bind cyclic oligoadenylate second messengers produced by the Type III interference complexes, the CARF proteins investigated therein are genetically encoded in association with the Type III systems producing the signals [6, 7], demonstrating *intra*-system regulation. The work here represents the first structural confirmation of *inter*-system coordination of Type I and Type III systems through the CARF/cA_4

interaction. Csa3a, a CARF family protein harboring a wHTH effector domain, is encoded within the Type I-A system while cA_4 is synthesized by the Type III interference complex. While previous functional studies of Csa3a in *S. islandicus* characterized its role in regulation of acquisition gene and CRISPR loci transcription, our functional characterization of *S. solfataricus* Csa3a reveal apparently higher complexity or control in its mechanism of transcriptional regulation as simply overexpressing Csa3a in *S. solfataricus* PH1-16 cells did not result in the same acquisition events as in *S. islandicus* [8, 9]. Seeded by initial binding sites identified by Liu et al., we identified a core binding motif, designated CAPPa, that is coincident with Csa3a in all but one sequenced Sulfolobales genomes and shows conserved synteny with Csa3a and Type I-A acquisition cassettes [8, 9]. Interestingly, Csa3a binds CAPPa in a nonspecific, cA_4-independent manner, suggesting its target specificity and/or regulation by cA_4 are more complex than originally hypothesized.

Future Studies

Investigating Potential Protein-Protein Interactions with Csa3a. One hypothesis that may explain the more complex regulatory mechanism exhibited by SsCsa3a would be a requirement of protein-protein interactions, particularly those that may involve the unstructured C-terminal tail it harbors that is not present in *S. islandicus* Csa3a. Of the potential interaction partners, those of TBP, TFB1, and TFB3 are the most intriguing as binding sites for these components of the

transcription initiation complex are found adjacent to the CAPPa motif associated with Csa3a. Whether these hypothesized interactions would be responsive to cA_4 binding by Csa3a or might alter Csa3a DNA-binding specificity remain only speculations. Unfortunately, initial attempts to test this theory were halted as *S. solfataricus* TBP, TFBI, and TFBIII expressed in *E. coli* were insoluble. Given pure, soluble TBP/TFBI/TFBIII constructs, protein-protein interactions could be probed via similar fluorescence polarization assays as used to characterize Csa3a DNA-binding activity. The possibility remains that, though Csa3a and CAPPa appear genetically linked, Csa3a may bind a different site in the genome. One rationale for additional complexity for SsCsa3a regulatory mechanisms as compared to SiCsa3a is the complexity of total CRISPR-Cas systems harbored in their respective genomes. While *S. islandicus* encodes a single Type I-A, two Type III-B systems, and two CRISPR loci, *S. solfataricus* encodes three Type I-A systems, three Type III-B, one Type III-D system and six CRISPR loci. As the complexity of the immune response grows with additional CRISPR-Cas systems, more finely tuned and complicated transcriptional regulation may be necessary.

Structural Characterization of Csa3b

The closely related Csa3b encoded near Csa3a in the *S. solfataricus* is another prime candidate for structural and functional characterization in the same manner. A core DNA binding motif for Csa3b was identified Q. Ye et al., which we refined, in unpublished work, to a core palindrome [10]. They described the role of

S. islandicus Csa3b in the regulation of Cmr interference genes, though our own attempts to structurally characterize *S. solfataricus* Csa3b were unsuccessful as expression in *E. coli* for generation of suitable quantities of protein was found to be toxic to the cells. Plasmids harboring the SsCsa3b gene were toxic to *E. coli* strains that did not have mechanisms to control leaky expression. Expression of SsCsa3b in yeast or other expression systems may be necessary to overcome its toxicity in *E. coli* and allow for characterization of the Csa3b/cA_4 complex as well as its cA_4-binding and DNA-binding ability.

Conclusions from Chapter Four

The research presented in Chapter Four began upon observation that the expression system we employed for a number of structural biology projects resulted in significantly increased yields of *Homo sapiens* Galectin-3 constructs compared to traditional methods of expression. Though our use of autoinduction to generate significant quantities of Galectins did not begin as a research project in its own right, we recognized the opportunity it presented as an approachable research project for undergraduate research. We combined this opportunity with the growing number of biochemistry students seeking work in a research setting for their required research credits. After recognizing that our method of autoinduction solved the problem of low Galectin yields, we moved the project forward with three goals, 1) generate protein stocks for biochemical assays, 2) characterize the increased efficiency of our methodology in comparison to that of traditional IPTG-

induction, and 3) incorporate a large number of undergraduate researchers into the work in a cooperative environment that was efficient on the both research and student progress fronts.

The first of these goals, generation of significant protein stocks was a carryover goal from before we decided to pursue this issue as a research project of its own. In this setting, though, we were able to express, purify, and store nearly a gram of protein stocks and test expression of eight galectins/galectin constructs (Galectin-3, Galectin-3 1-163, Galectin-3 1-111, Galectin-3 63-250, Galectin-3 112-250, Galectin-7, Galectin-8, and Galectin-9). While they did not appear in the published material in Chapter Four, we also incorporated proteins for a variety of other projects into the schedule/workflow including *S. solfataricus* Csa3a, *S. solfataricus* Csa3b, *S. solfataricus* Csa3c, *S. solfataricus* Hmg-CoA reductase, *S. islandicus* Csa3a, *Homo sapiens* DYRK1A kinase, and others. Many of these proteins expressed well enough for initial investigations while others did not express well enough for use in this setting.

We found in early experiments that the yields of Galectin-3 112-250, the carbohydrate recognition domain (CRD), were nearly an order of magnitude higher in our autoinduction method than for the traditional IPTG-induction methods used by most labs. Additionally, attempts to increase yields in IPTG-induction by increasing IPTG concentrations were insignificant until we increased IPTG concentrations to 5 mM, 50x the concentration used in our standard expression conditions. More surprising was our uninduced control culture from which we

purified equivalent quantities of Galectin-3 CRD as from cultures induced with up to 2 mM IPTG. This observation suggested that all expression of Galectin-3 CRD that we observed at IPTG concentrations below 2 mM IPTG was a result of leaky expression and induction did not occur in these cultures. We then characterized the binding affinity of Galectin-3 CRD for IPTG and found a K_d of 64 ± 5 μM, which, when put into perspective with a cellular Galectin-3 CRD concentration of ~240 μM from leaky expression alone, suggests induction at typical IPTG concentrations may not occur as Galectin-3 CRD likely binds and sequesters significant quantities of it. When the same cells and vectors are used for expression of Galectin-3 CRD in autoinduction, leaky expression may be more limited due to the presence of glucose and it's inhibitory effects on the import of lactose and inhibition of T7 polymerase transcription that is under control of the lacUV5 promoter.

Finally, the research performed in Chapter Four was done in a setting to fulfill the research credit requirements for fourteen undergraduate students and included work from two high school students. Modeled after a curriculum-based undergraduate research experience (CURE), the system included around 8-10 hours per week of hands-on research experiment time for each student in addition to 1-2 hours per week of classroom-based instruction. While the time involvement per student was standard for most undergraduate researchers with a full class schedule, this setting was unique in allowing students to take part in all aspects of the research project from start to finish since progress was made far more quickly than usual undergraduate projects because of the combined efforts of sixteen

students. Though initial scheduling, setup, and instruction was time consuming, students were soon able to teach their peers lab skills without significant supervisor input, and this, combined with coordination of experimental tasks, allowed teams of 2-6 students to complete experiments in shorter time periods than a lone undergraduate researcher is capable of.

This method used to bypass a putative inhibitory feedback mechanism limiting expression of Galectins, and likely many other carbohydrate-binding proteins, will significantly reduce the cost, time, and effort required for their characterization. Further, this method is approachable even to undergraduate researchers with no prior research experience.

References

1. Haft, D., et al., *A Guild of 45 CRISPR-Associated (Cas) Protien Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.* Plos Computational Biology, 2005. **1**(6): p. 474-483.

2. Makarova, K.S., et al., *Evolution and classification of the CRISPR-Cas systems.* Nat Rev Microbiol, 2011. **9**(6): p. 467-77.

3. Makarova, K.S. and E.V. Koonin, *Annotation and Classification of CRISPR-Cas Systems.* Methods Mol Biol, 2015. **1311**: p. 47-75.

4. Makarova, K.S., et al., *An updated evolutionary classification of CRISPR-Cas systems.* Nat Rev Microbiol, 2015. **13**(11): p. 722-36.

5. 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.

6. Kazlauskiene, M., et al., *A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems.* Science, 2017. **357**(6351): p. 605-609.

7. Niewoehner, O., et al., *Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers.* Nature, 2017. **548**(7669): p. 543-548.

8. Liu, T., et al., *Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition.* Nucleic Acids Res, 2015. **43**(2): p. 1044-55.

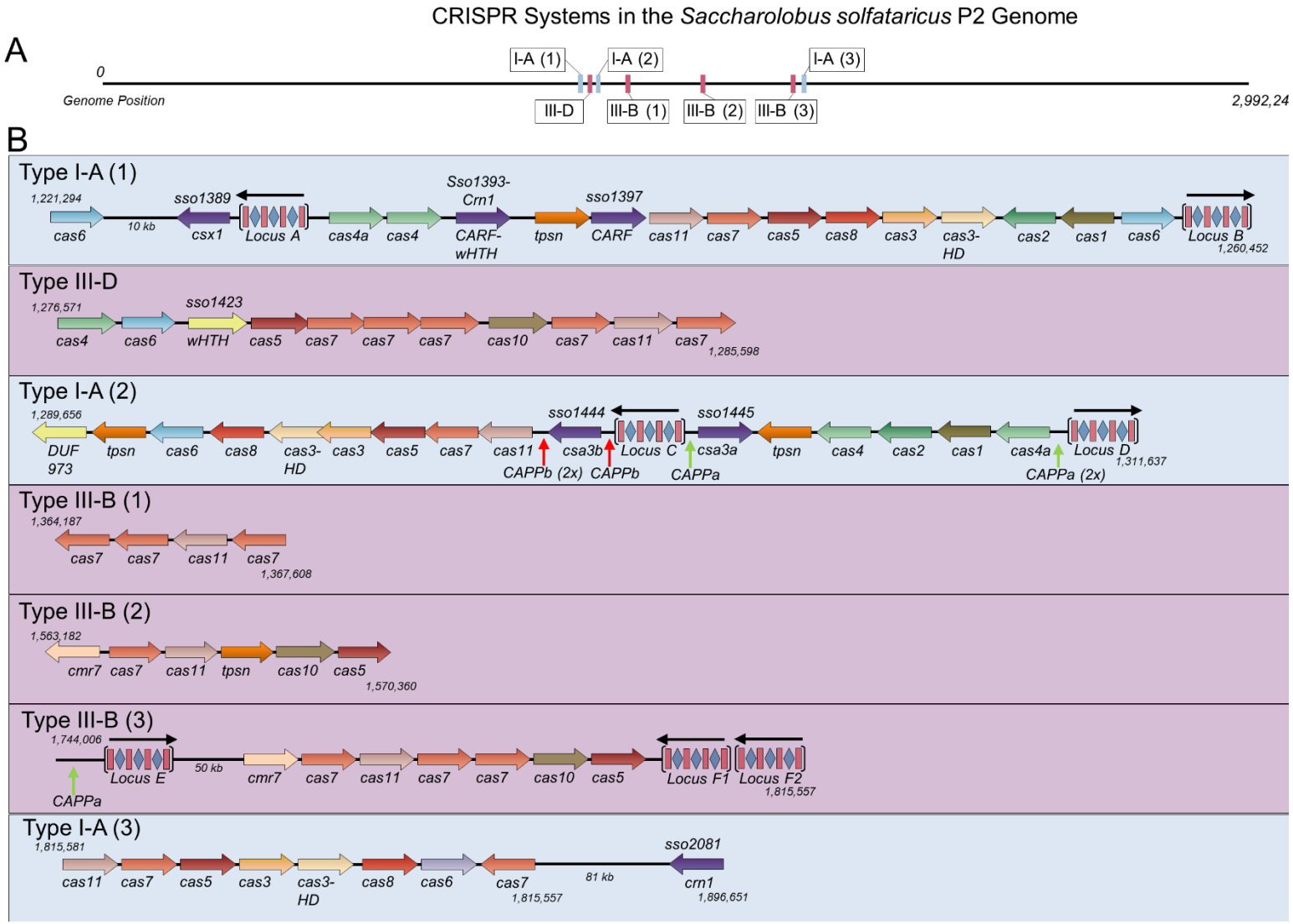
9. Liu, T., et al., *Coupling transcriptional activation of CRISPR–Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus.* Nucleic Acids Research, 2017. **45**(15): p. 8978-8992.

10. Ye, Q., et al., *CRISPR-Associated Factor Csa3b Regulates CRISPR Adaptation and Cmr-Mediated RNA Interference in Sulfolobus islandicus.* Frontiers in Microbiology, 2020. **11**.

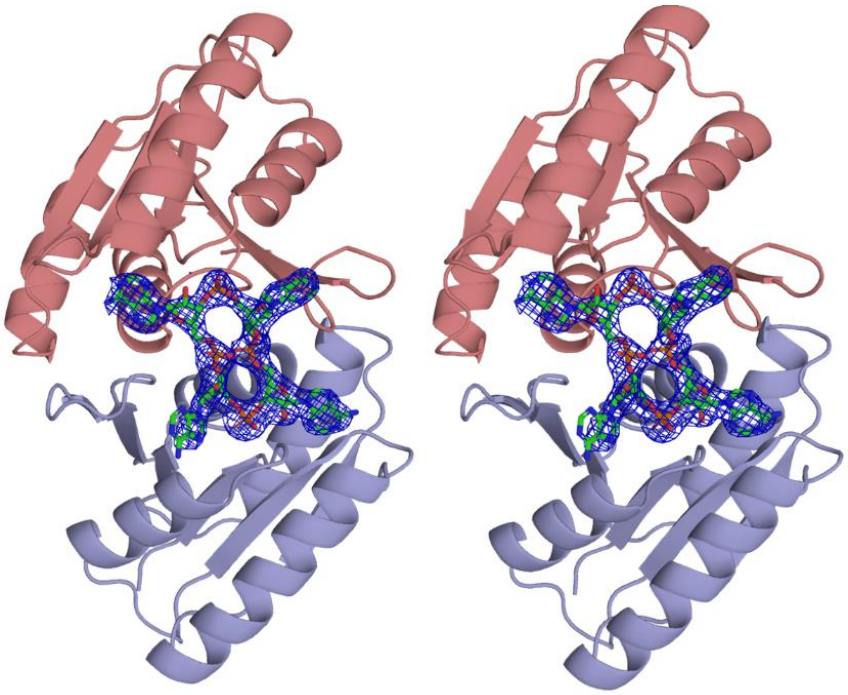
APPENDICES

APPENDIX A

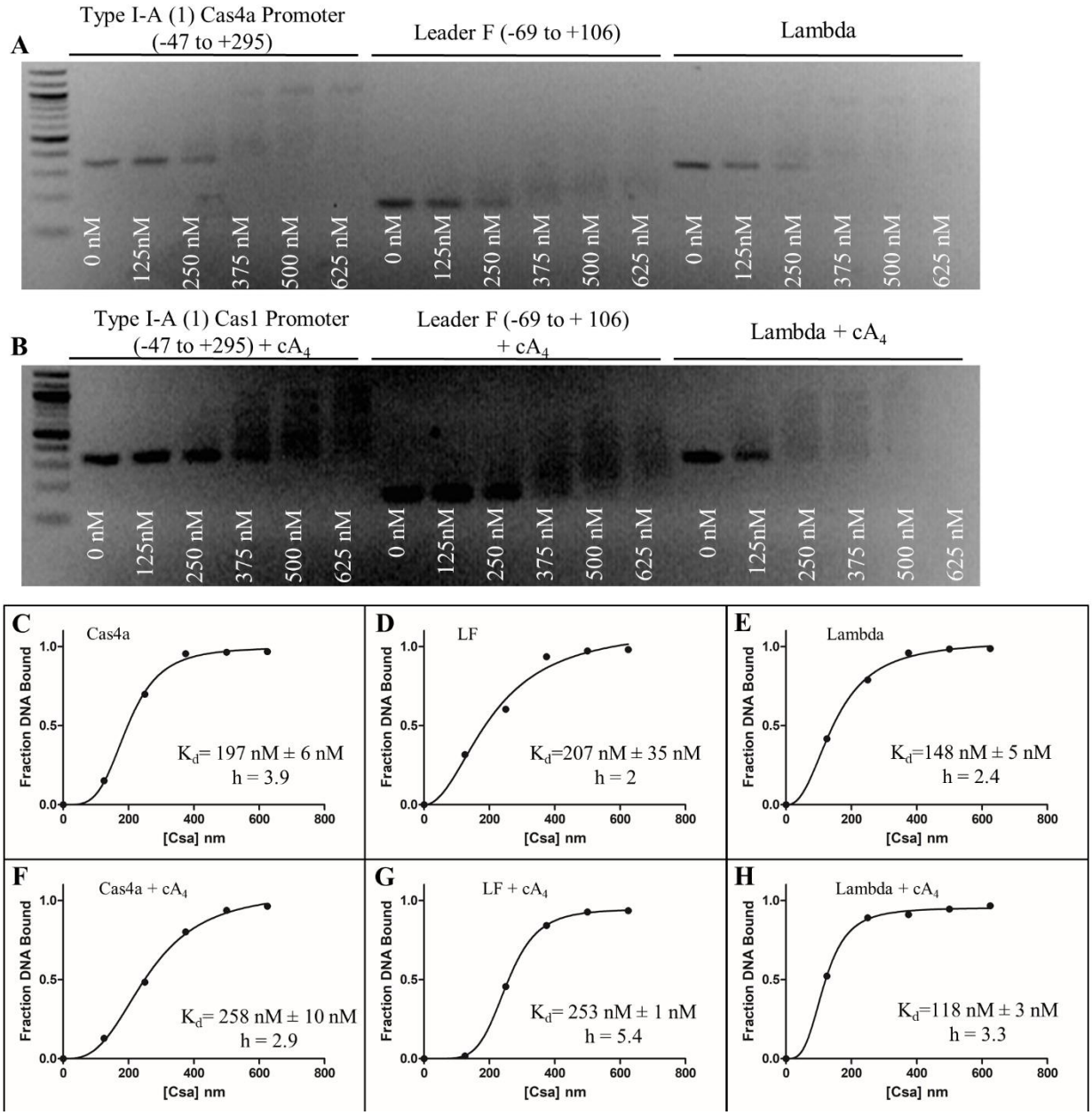
Supporting Online Materials For: Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response


CRISPR Systems in the Saccharolobus solfataricus P2 Genome
A
0
Genome Position
2,992,24
I-A (1)
I-A (2)
I-A (3)
III-D
III-B (1)
III-B (2)
III-B (3)
B
Type I-A (1)
1,221,294
cas6
10 kb
sso1389
csx1
Locus A
cas4a
cas4
Sso1393-Crn1
CARF-wHTH
tpsn
sso1397
CARF
cas11
cas7
cas5
cas8
cas3
cas3-HD
cas2
cas1
cas6
Locus B
1,260,452
Type III-D
1,276,571
cas4
cas6
sso1423
wHTH
cas5
cas7
cas7
cas7
cas10
cas7
cas11
cas7
1,285,598
Type I-A (2)
1,289,656
DUF 973
tpsn
cas6
cas8
cas3-HD
cas3
cas5
cas7
cas11
CAPPb (2x)
csa3b
CAPPb
sso1444
Locus C
sso1445
csa3a
CAPPa
tpsn
cas4
cas2
cas1
cas4a
CAPPa (2x)
Locus D
1,311,637
Type III-B (1)
1,364,187
cas7
cas7
cas11
cas7
1,367,608
Type III-B (2)
1,563,182
cmr7
cas7
cas11
tpsn
cas10
cas5
1,570,360
Type III-B (3)
1,744,006
Locus E
CAPPa
50 kb
cmr7
cas7
cas11
cas7
cas7
cas10
cas5
Locus F1
Locus F2
1,815,557
Type I-A (3)
1,815,581
cas11
cas7
cas5
cas3
cas3-HD
cas8
cas6
cas7
1,815,557
81 kb
sso2081
crn1
1,896,651

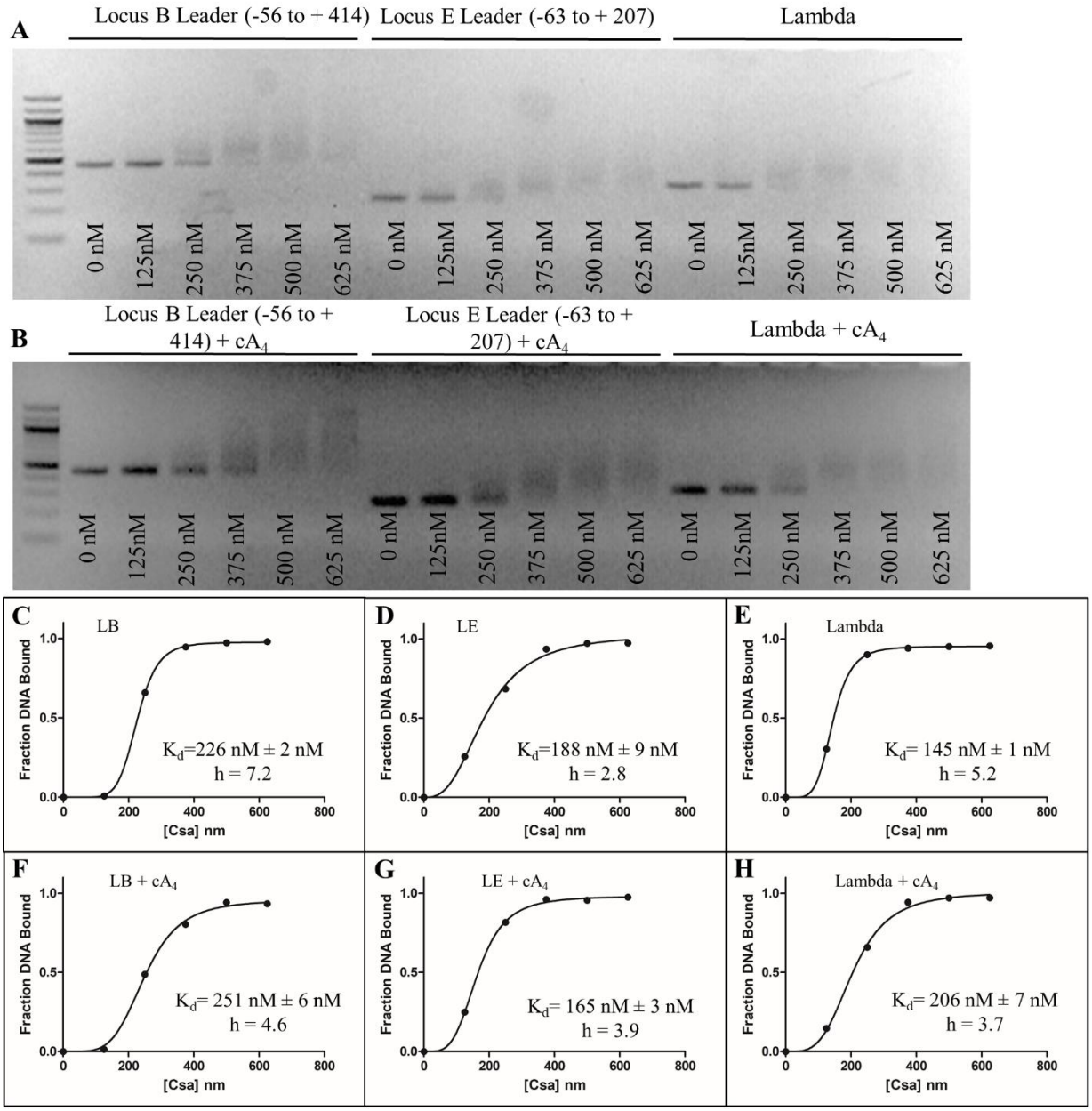
Supplementary Figure S1. Overview of CRISPR/Cas elements in S. solfataricus P2. **A)** General locations of the CRISPR/Cas clusters are represented by blue (Type I clusters) and red (Type III) clusters) bars on a black line representing the ~3 Mbp genome. **B**) Genes are colored based on the encoded gene. The acquisition genes, *cas1, cas2, cas4, cas4a* are in shades of green and *cas6* (crRNA maturation) is in blue. The interference genes are in shades of red to yellow. CARF domain genes are violet. Non-*cas* genes or genes of unknown function are colored bright yellow, and transposons are colored bright orange. CRISPR loci are depicted as repeated red bars (repeats) and blue diamonds (spacers). The positions of CAPPa are indicated by green vertical arrows. The locations of an unrelated binding motif reported for Csa3b [1], which we denote CAPPb, are also indicated as vertical red arrows. Italic numbers at either end of the clusters denote genome positions.



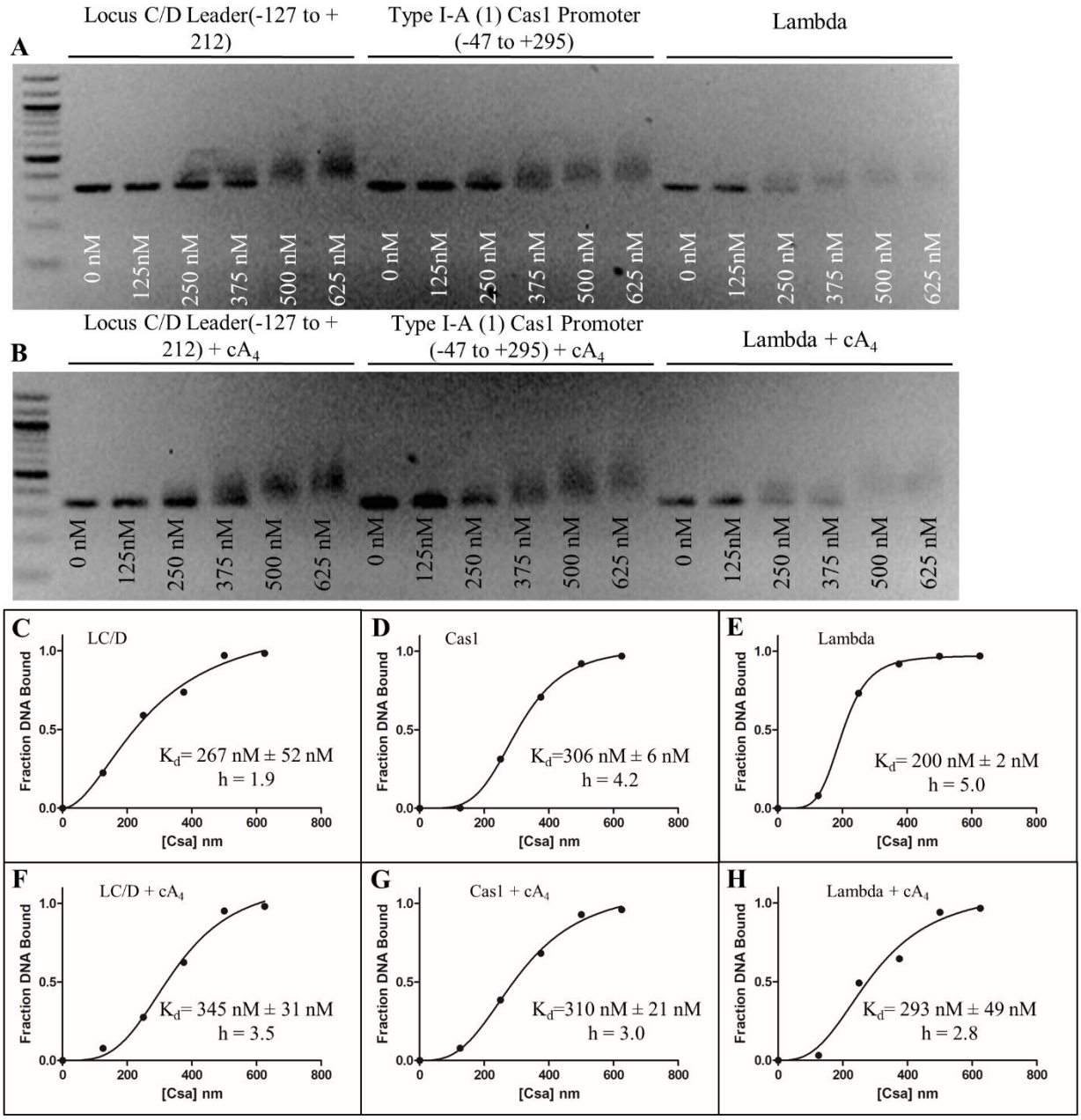
Supplemental Figure S2. cA4 omit map. Ribbon diagram of the Csa3 dimer bound to cA_4. Chain A is blue and chain B is red. cA_4 is in a stick representation with C atoms in green. The $2F_o$-F_c electron density omit map contoured at 1.0σ (blue) was generated with a single round of refinement with Csa3 without cA_4.



Supplemental Figure S3. Nonspecific binding to the I-A (1) cas4a promoter and CRISPER locus F leader. **A)** Csa3 binds all fragments with ~200 nm affinity, including Lambda phage DNA indicating a nonspecific interaction in the absence of cA_4. **B)** Addition of cA_4 does not affect binding affinity or specificity. EMSA experiments were performed with 18 nM DNA and increasing [Csa3] dimer (denoted below the lanes). EMSA fragments are detailed in Supplemental Table S4. Lambda phage DNA was used as a negative control. **C-H.** Graphs generated in GraphPad Prism display curves for specific binding with a Hill coefficient (h) fit to binding data generated using ImageJ.



Supplemental Figure S4. Nonspecific binding to CRISPR loci B and E leaders. **A)** Csa3 binds all fragments with ~200 nm affinity, including Lambda phage DNA indicating a nonspecific interaction in the absence of cA_4. **B)** Addition of cA_4 does not affect binding affinity or specificity. EMSA experiments were performed with 18 nM DNA and increasing [Csa3] dimer (denoted below the lanes). EMSA fragments are detailed in Supplemental Table S4. Lambda phage DNA was used as a negative control. **C-H.** Graphs generated in GraphPad Prism display curves for specific binding with a Hill coefficient (h) fit to binding data generated using ImageJ.



Supplemental Figure S5. Nonspecific binding to the CRISPR locus F leader and I-A(1) cas1 promoter. **A**) Csa3 binds all fragments with ~200-300 nm affinity, including Lambda phage DNA indicating a nonspecific interaction in the absence of cA_4. **B**) Addition of cA_4 does not affect binding affinity or specificity. EMSA experiments were performed with 18 nM DNA and increasing [Csa3] dimer (denoted below the lanes). EMSA fragments are detailed in Supplemental Table S4. Lambda phage DNA was used as a negative control. **C-H.** Graphs generated in GraphPad Prism display curves for specific binding with a Hill coefficient (h) fit to binding data generated using ImageJ.

Supplemental Tables

Supplemental Table S1 – ITC Values

[cA4] (μM)	K_D (μM)*	ΔH_{25^oC} (Cal/mol)*	ΔS_{25^oC} (Cal/mol/deg)*	N*†
250	1.09 (0.10)	13,490 (390)	72.5 (1.2)	0.546 (0.028)
500	1.48 (0.19)	13,380 (470)	71.6 (1.3)	0.608 (0.017)

*Duplicate ITC runs were averaged, with the range indicated.
†Stoichiometry of ligand:protein interaction relative to the Csa3a dimer.

Supplemental Table S2 - PCR Primers For EMSA Fragments

Fragment Name	Annotation	Size (bp)	Genome Location	Annealing Temp. °C	Primers (5' to 3') - Forward (top), Reverse (bottom)
Cas4a (II)	Promoter region for Cas4a (Group 2 acquisition cassette)	350	1305123-1305472	63	ATAACGTCTAGGCTTTATCGGAGG CGTTTGATGTTAGTTTCTTAGGAGGTA
Cas4a (I)	Promoter region for Cas4a/Cas4 (Group 1 acquisition cassette)	342	1240163-1240504	63	CCGCAACTAACCCGTTCC TTTATCCTCTTTGAAAGCAACAGC
Cas1	Promoter region for Cas1/Cas2 (Group 1 acquisition cassette)	350	1253317-1253666	63	GAATGGAGTTCTGTAGGGGCAA TAATCGAAATGTAAGTGTGTAAATCACAG
LB	Promoter region directly upstream of CRISPR Leader B	470	1260396-1260865	63	GATATTATACCGGTTGAGCCTGC CAGCCTTAGTGAAATTAGGAAATATTACG
LC/D	Promoter region directly upstream of CRISPR Leaders C/D	339	1305422-1305760	63	TGTCCCGTTTTTGTAAGTGGG TCGATAGAAAGATTTAAATATGCGGAATAC
LE	Promoter region directly upstream of CRISPR Leader E	270	1743799-1744068	64	GGGAAAGAGTTCCCCCGTA ACGTCGCTGCCAATTTTCC
LF	Promoter region directly upstream of CRISPR Leader F	175	1811222-1811396	62	CTGTTAGTAAGCCCCTTGCTAC AGAGAATACTTGTTACCTAACTGGT
Lambda	Control Fragment from Lambda phage genome	350	24398-24747	62	TAGGAATTGGTTAGCAAGTTACTACC TTGCTAACAGGTATCGTTTGGAG

Supplemental Table S2. PCR primers used to generate EMSA fragments. The fragment name denotes the name used to reference the indicated fragment in experimental procedures and results. The genome location indicates the location of the amplified fragment within the genome of *S. solfataricus* P2 for all fragments besides Lambda-350 which was generated from the Lambda phage genome.

Supplemental Table S3 - Fluorescence Polarization DNA Constructs

Fragment Name	Size (bp)	*S. solfataricus* P2 Genome Location	Annealing Temp. °C	ssDNA Oligonucleotides or Primers (5’ to 3’)
20bp 5'-FAM I-A(2) Cas4a CAPPa	20	1305298-1305317	N/A	/56-FAM/TTCCCGAATTTACTAGGGAT AAGGGCTTAAATGATCCCTA
28bp 5'-FAM Idealized CAPPa	28	N/A	N/A	/56-FAM/GTTAATCCCGAATTAATTCGGGATTAAC (self-complimentary)
75bp 5'-FAM I-A(2) Cas4a CAPPa	75	1305298-1305372	61	/56-FAM/TTCCCGAATTTACTAGGGAT TCCCTAGTTTAATTATTAATCTTTATATAGAGATGA

Supplemental Table S4 - Acquisition Assay PCR Primers

Fragment Name	Size (bp)	*S. solfataricus* P1 Genome Location	Annealing Temp. °C	Primers (5’ to 3’) - Forward (top), Reverse (bottom)
Locus A	147	1321505-1321651	57.5	ACTAAAATTGGTCGCATGAAGAGTA ACTTAGCAGAATATCTACTGAAATATCTGC
Locus C	147	1384343-1384489	57.6	AACACTCCCGTACCAATTTCTTATT CTAACACAATCCGATTAAAGTATATCCTCA
Locus D	170	1392982-1393151	62	CAGAGAAAAGCTTATAAATAACTAAGGAGA GGTCGACAATCCCGTAGATC
Locus E	152	1817367-1817518	56.3	GGAAAACCCGGGAATAGTAAAAA TGCCAATTTTCCGCCAATTC
Locus F	188	1883103-1883290	54.8	CCTGATTTATTCAATCCAGAGAATTT GGCAATAGTGAATGGTGTAGTTATA
Locus F2	156	1884684-1884839	57.7	TAGTAAGCCCCTTGCTACAACT CCTAACTGGTTGAGCAAAAACTTTAAATAT

Supplemental Table S5 – Acquisition Cassette Synteny

Strain																			
A. brierleyi DSM 1651	*csa3b*	**-II-**	< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >			
A. hospitalis W1		< *csa3b*	< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2* >	CAPPa	*cas4* >	< *csa3a*	CAPPa	Locus >			
A. sulfidivorans JP7		< *csa3b*	< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >			
M. prunae Ron 12		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula ARS120-1		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula ARS120-2		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula ARS50-1		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula ARS50-2		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula DSM 5348		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. sedula SARC-M1		< *csa3b*	< Locus	CAPPa	CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	Csa3a >	CAPPa	Locus >			
M. tengchongensis Ric-A			< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >			
S. solfataricus P1			< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. solfataricus P2			< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. solfataricus POZ149			< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. sp. E-5-1-F			< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	**-II-**	*csa3b*	
S. sp. S-194			< Locus			CAPPa	< *csa3a*	CAPPa	*cas4a* >	*cas1* >	*cas2* >		*cas4* >		CAPPa	Locus >			
S. islandicus HVE10-4		< *csa3b*	< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2*/CAPPa		*cas4* >	< *csa3a*	CAPPa	Locus >			
S. islandicus L.S.2.15			< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. islandicus LAL14/1			< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2*/CAPPa		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. islandicus M.14.25		< *csa3b*	< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >			
S. islandicus M.16.27			< Locus		CAPPa	CAPPa			*cas4a* >	*cas1* >	*cas2* >		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. islandicus REY15A			< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2*/CAPPa		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. islandicus Y.G.57.14		< *csa3b*	< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2*/CAPPa		*cas4* >	< *csa3a*	CAPPa	Locus >			
S. islandicus YN.15.51			< Locus			CAPPa			*cas4a* >	*cas1* >	*cas2*/CAPPa		*cas4* >	< *csa3a*	CAPPa	Locus >	*csa3b* >		
S. tokodaii 7			< Locus		CAPPa	CAPPa	< *csa3a*	CAPPa	*cas4a* >	*cas1* >	*cas2* >		*cas4* >			< Locus		CAPPa	CAPPa
S. ohwakuensis TA-1	*csa3b*	**-II-**	< Locus		CAPPa	CAPPa	< *csa3a*	CAPPa	*cas4a* >	*cas1* >	*cas2* >		*cas4* >			< Locus		CAPPa	CAPPa

*single nucleotide mutation

Supplemental Table S5. Acquisition cassette synteny for Sulfolobales containing Csa3a. The csa3a gene is generally associated with the acquisition cassette (green genes), and a CAPPa sequence is also present immediately upstream of the cas4a gene. Gene and locus transcript direction are indicated by the adjacent arrows. Gene positions are relative, and transposases or proteins of unknown function are not shown. The -//- symbolizes large genomic distance between genes.

Supplemental Table S6 – Correlation of Csa3a and CAPPa

Strain	aCas	Csa3a	CAPPa
Acidianus ambivalens LEI 10	-	-	-
Acidianus brierleyi DSM 1651	+	+	+
Acidianus hospitalis W1	+	+	+
Acidianus manzaensis YN-25	-	-	-
Acidianus sulfidivorans JP7	+	+	+
Metallosphaera cuprina Ar-4	-	-	-
Metallosphaera hakonensis JCM 8857	-	-	-
Metallosphaera prunae Ron 12	+	+	+
Metallosphaera sedula ARS120-1	+	+	+
Metallosphaera sedula ARS120-2	+	+	+
Metallosphaera sedula ARS50-1	+	+	+
Metallosphaera sedula ARS50-2	+	+	+
Metallosphaera sedula DSM 5348	+	+	+
Metallosphaera sedula SARC-M1	+	+	+
Metallosphaera tengchongensis Ric-A	+	+	+
Saccharolobus solfataricus P1	+	+	+
Saccharolobus solfataricus P2	+	+	+
Saccharolobus solfataricus POZ149	+	+	+
Saccharolobus solfataricus 98/2*	+	-	+
Stygiolobus azoricus FC6	-	-	-
Sulfodiicoccus acidiphilus HS-1	+	-	-
Sulfolobus acidocaldarius DG1	-	-	-
Sulfolobus acidocaldarius DSM 639	-	-	-
Sulfolobus acidocaldarius N8	-	-	-
Sulfolobus acidocaldarius Ron12/I	-	-	-
Sulfolobus acidocaldarius SUSAZ	-	-	-
Sulfolobus acidocaldarius Y14 13-1	-	-	-
Sulfolobus acidocaldarius Y14 16-22	-	-	-
Sulfolobus acidocaldarius Y14 18-5	-	-	-
Sulfolobus acidocaldarius Y14 20-20	-	-	-
Sulfolobus islandicus HVE10/4	+	+	+
Sulfolobus islandicus L.S.2.15	+	+	+
Sulfolobus islandicus LAL14/1	+	+	+
Sulfolobus islandicus M.14.25	+	+	+
Sulfolobus islandicus M.16.27	+	+	+
Sulfolobus islandicus M.16.4	+	**	+
Sulfolobus islandicus REY15A	+	+	+
Sulfolobus islandicus Y.G.57.14	+	+	+
Sulfolobus islandicus Y.N.15.51	+	+	+
Sulfolobus sp. A20	-	-	-
Sulfolobus sp. E11-6	-	-	-
Sulfolobus sp. E5-1-F	+	+	+
Sulfolobus sp. S-194	+	+	+
Sulfolobus tokodaii 7	+	+	+
Sulfuracidifex tepidarius IC-006	-	-	-
Sulfuracidifex tepidarius IC-007	-	-	-
Sulfurisphaera ohwakuensis TA-1	+	+	+

Supplemental Table S6. Correlation of Csa3a and CAPPa. Csa3 presence or absence correlates strongly with the presence or absence of CAPPa in the acquisition cassette (aCas). A "+" denotes the presence of the indicated element, "-" denotes absence, "*" indicates combined parent and daughter strains and "**" indicates viral insertion into csa3a gene. csa3a is present in all sequenced genomes with CAPPa except those derived from *S. solfataricus* 98/2.

Supplemental Table S7 - RNA Conformers for cA_4*

Nucleotide	δ,δ,γ	Name	Suite	δ-1	ε	ζ	α	β	γ	δ
A1	22m	2o	2'emmtm2'	158.08	-145.90	50.51	-	-119.33	-71.42	156.24
A2	22p	6p	2'epptp2'	156.24	-151.05	125.27	61.82	167.29	49.25	155.06
A3	22m	2o	2'emmtm2'	155.06	-146.90	75.04	70.37	-165.48	-66.01	153.03
A4	22p	6p	2'epptp2'	153.03	-	-	74.01	-172.64	55	158.08

*For RNA backbone conformer nomenclature see Richardson et al. [2].

References

1. He, F., et al., *CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b.* Nucleic Acids Res, 2017. **45**(4): p. 1902-1913.

2. Richardson, J.S., et al., *RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution).* RNA, 2008. **14**(3): p. 465-81.

APPENDIX B

Supplementary Data: A CURE for High Level Galectin Expression

Table S1. PCR Primers for Galectin-3 and Galectin-3 CRD

Construct	Annealing Temp. ℃	Primers (5' to 3') - Forward (top), Reverse (bottom)
Galectin-3 (Full Construct)	70	ACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAA CCATGCACCATCACCATCACCATGCAGACAATTTTTCGCTCC CGCACCACTTTGTACAAGAAAGCTGGGTC TTATATCATGGT ATATGAAGCACTGG
Galectin-3 CRD	70	ACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAA CCATGCACCATCACCATCACCATGGGCCACTGATTGTG CGCACCACTTTGTACAAGAAAGCTGGGTC TTATATCATGGT ATATGAAGCACTGG

Galectin-7

ATGCATCACCATCATCATCATTCAAACGTACCTCACAAGTCGTCTCTGCCTGAG
GGTATTCGTCCCGGAACGGTCCTGCGAATACGTGGTCTGGTTCCGCCGAATGC
GTCACGATTTCACGTCAATCTGTTGTGTGGGGAAGAACAGGGCTCCGATGCAG
CTCTGCATTTCAACCCTCGTCTGGACACCTCGGAGGTAGTATTTAATAGCAAAG
AGCAGGGGTCTTGGGGGCGTGAAGAGCGCGGGCCGGGTGTGCCGTTTCAGCG
TGGTCAGCCTTTTGAAGTGCTGATCATTGCGTCTGATGACGGATTCAAGGCGG
TTGTGGGGGACGCACAATATCATCATTTTAGACACCGTTTACCTCTGGCCCGTG
TGCGTTTAGTAGAAGTTGGCGGAGATGTCCAATTAGATTCGGTGCGTATTTTTT
AA

Galectin-8

ATGCATCATCATCATCATCACATGATGCTGTCATTGAATAATCTGCAAAATATC
ATATACAATCCGGTCATTCCATATGTCGGTACAATCCCAGATCAACTGGATCCG
GGTACCCTGATTGTGATTTGCGGTCATGTACCATCAGATGCAGATCGCTTTCAG
GTAGACCTCCAGAACGGTTCGTCAGTTAAACCCCGGGCCGATGTGGCTTTTCA
CTTTAATCCTCGCTTCAAGCGTGCGGGTTGTATTGTCTGCAACACTTTGATAAA
CGAAAAATGGGGACGTGAAGAAATTACGTATGATACACCCTTCAAACGCGAAA
AATCCTTTGAAATCGTCATCATGGTTCTTAAAGATAAATTTCAGGTGGCCGTCA
ACGGGAAGCATACGTTGCTGTATGGACATCGTATTGGACCAGAAAAGATCGAT
ACCTTGGGTATCTACGGTAAGGTTAATATCCACTCAATTGGATTTTCTTTCTCC
TCAGACCTGCAATCTACGCAAGCGTCAAGCCTGGAACTGACCGAAATAAGCCG
TGAAAACGTTCCTAAATCTGGAACCCCGCAACTCAGTTTGCCGTTTGCTGCTCG
CCTGAATACCCCCATGGGTCCAGGTCGAACGGTTGTCGTGAAAGGTGAAGTTA
ATGCTAACGCTAAGTCTTTTAATGTTGATCTGTTGGCAGGTAAAAGCAAGGATA
TAGCTCTGCATCTGAACCCTCGTCTGAATATTAAGGCCTTCGTGCGCAACTCGT
TTCTTCAAGAGTCTTGGGGTGAAGAAGAACGCAACATAACCAGCTTCCCATTTT
CACCAGGCATGTACTTCGAGATGATCATCTATTGTGATGTTCGGGAATTCAAGG

TGGCGGTCAATGGTGTGCACTCTCTTGAGTACAAGCATCGTTTTAAGGAATTAA
GTAGCATTGATACCCTGGAGATCAACGGCGACATCCATCTGCTGGAGGTGCGT
AGTTGGTAA

Fig. S1. Nucleotide sequences for Galectin-7 and Galectin-8

Table S2. Measured extinction coefficients

Construct	Molecular weight (g/mol)	Extinction coefficient	
Galectin-3	26,975	16,454	A_{280} = 0.61 @ 1 mg/mL, 37.1 μM
Galectin-3 CRD	16,594	9127	A_{280} = 0.55 @ 1 mg/mL, 60.3 μM
Galectin-7	15,899	8744	A_{280} = 0.55 @1 mg/mL, 62.9 μM (monomer), 31.45 μM (dimer)
Galectin-8	36,362	17,454	A_{280} = 0.48 @ 1 mg/mL, 27.5 μM

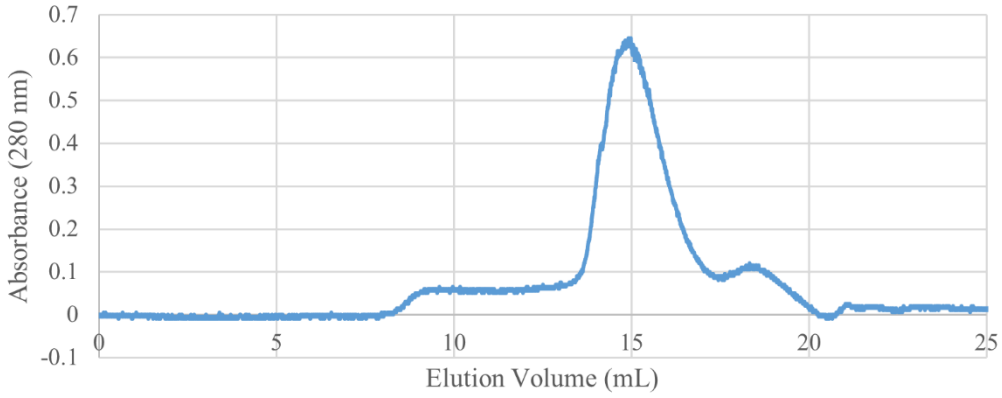


Fig. S2. Galectin-3 size exclusion chromatography. When applied to a calibrated Superdex 75 10/300 GL column (Cytiva, Marlborough, MA, USA) equilibrated with PBS buffer, purified *Homo sapiens* Galectin-3 elutes at ~14.9 mL, corresponding to a molecular weight of 20 kDa. With an expected monomeric mass of 26.9 kDa, it appears movement of full-length Gal-3 through the column is also retarded, again perhaps due to the dextran matrix. Regardless, as for Gal3-CRD, this results in highly pure Gal-3.

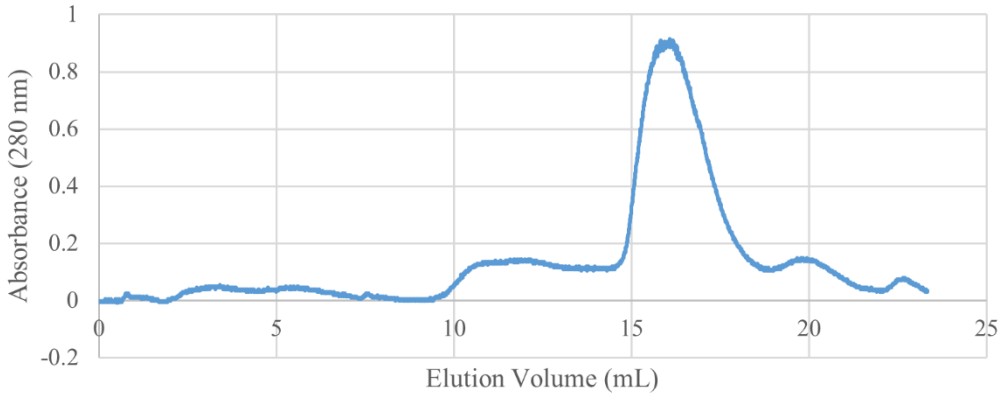


Fig. S3. Galectin-3 CRD size exclusion chromatography. When applied to a calibrated Superdex 75 10/300 GL column (Cytiva, Marlborough, MA, USA) equilibrated with PBS buffer, purified *Homo sapiens* Galectin-3 CRD elutes at ~16 mL, corresponding to a molecular weight of 11 kDa. With an expected monomeric mass of 16.5 kDa, it appears Gal-3 CRD is interacting with the column, perhaps the dextran matrix, retarding its movement. Regardless of the reason, this gives highly pure Gal-3 CRD following size exclusion chromatography.

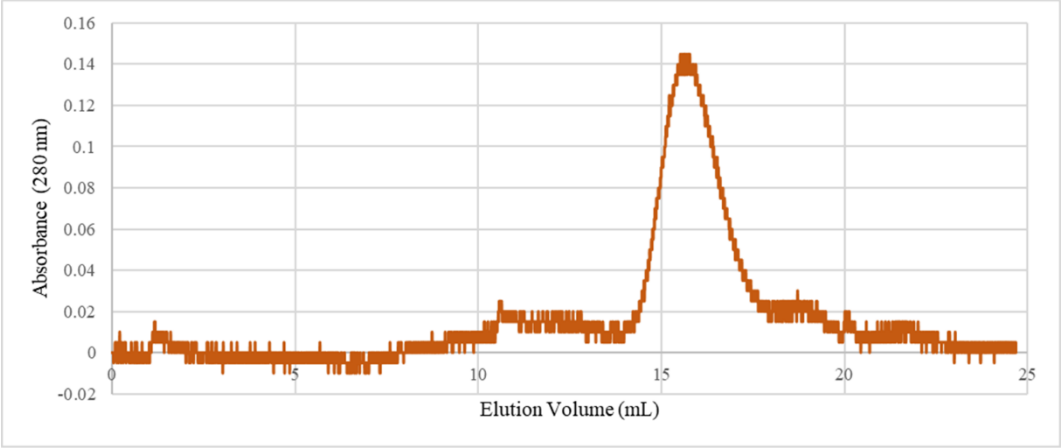


Fig. S4. Galectin 7 size exclusion chromatography. When applied to a calibrated Superdex 75 10/300 GL column (Cytiva, Marlborough, MA, USA) equilibrated with PBS buffer, purified *Homo sapiens* Galectin 7 elutes at ~15.75 mL, corresponding to a molecular weight of 13 kDa. With an expected subunit mass of 15.9 kDa, and 32 kDa for the anticipated dimer, movement of Galectin-7 through the column is also clearly retarded. Like the Gal-3 constructs, this may be due to an interaction with the dextran matrix. Regardless, like the Gal-3 constructs, the result is extremely pure Gal-7.

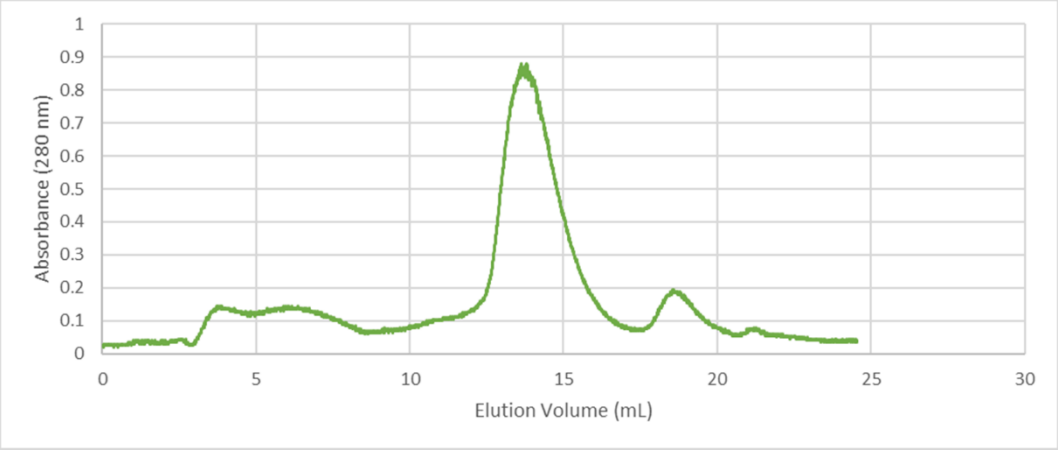


Fig. S5. Galectin-8 elutes as a monomer via size exclusion chromatography. When applied to a calibrated Superdex 75 10/300 GL column (Cytiva, Marlborough, MA, USA) equilibrated with PBS buffer, purified *Homo sapiens* Galectin 8 elutes at ~13.7 mL, corresponding to a molecular weight of 38 kDa. With an expected monomeric mass of 36.6 kDa, Galectin 8 is likely present as a monomer.

Gel-Based MS/MS Analysis – Orbitrap Fusion (Performed by the IDEA National Resource for Quantitative Proteomics at the University of Arkansas Medical School)

Each SDS-PAGE gel band was subjected to in-gel trypsin digestion as follows. Gel segments were destained in 50% methanol (Fisher), 50 mM ammonium bicarbonate (Sigma-Aldrich), followed by reduction in 10 mM tris(2-carboxyethyl)phosphine (Pierce) and alkylation in 50 mM iodoacetamide (Sigma-Aldrich). Gel slices were then dehydrated in acetonitrile (Fisher), followed by addition of 100 ng porcine sequencing grade modified trypsin (Promega) in 50 mM ammonium bicarbonate (Sigma-Aldrich) and incubation at 37 °C for 12-16 hours. Peptide products were then acidified in 0.1% formic acid (Pierce). Tryptic peptides were separated by reverse phase XSelect CSH C18 2.5 um resin (Waters) on an in-line 150 x 0.075 mm column using a nanoAcquity UPLC system (Waters). Peptides were

eluted using a 60 min gradient from 98:2 to 65:35 buffer A:B ratio. [Buffer A = 0.1% formic acid, 0.5% acetonitrile; buffer B = 0.1% formic acid, 99.9% acetonitrile.] Eluted peptides were ionized by electrospray (2.4 kV) followed by MS/MS analysis using higher-energy collisional dissociation (HCD) on an Orbitrap Fusion Tribrid mass spectrometer (Thermo) in top-speed data-dependent mode. MS data were acquired using the FTMS analyzer in profile mode at a resolution of 240,000 over a range of 375 to 1500 m/z. Following HCD activation, MS/MS data were acquired using the ion trap analyzer in centroid mode and normal mass range with precursor mass-dependent normalized collision energy between 28.0 and 31.0. Proteins were identified by database search using Mascot (Matrix Science) with a parent ion tolerance of 3 ppm and a fragment ion tolerance of 0.5 Da. Scaffold (Proteome Software) was used to verify MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established with less than 1.0% false discovery by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established with less than 1.0% false discovery and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm [1].

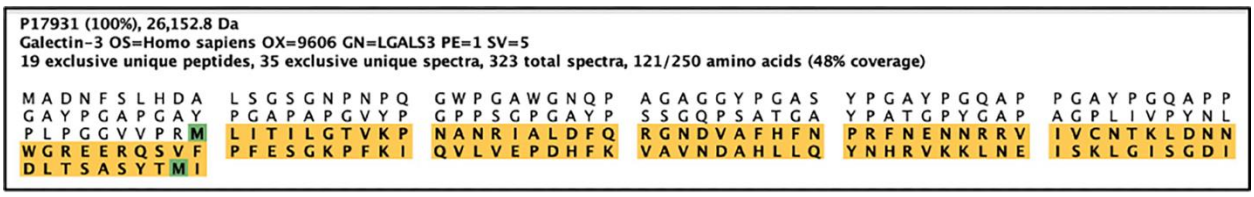


Fig. S6. Gal-3 CRD MS/MS results



Fig. S7. Galectin-7 MS/MS results

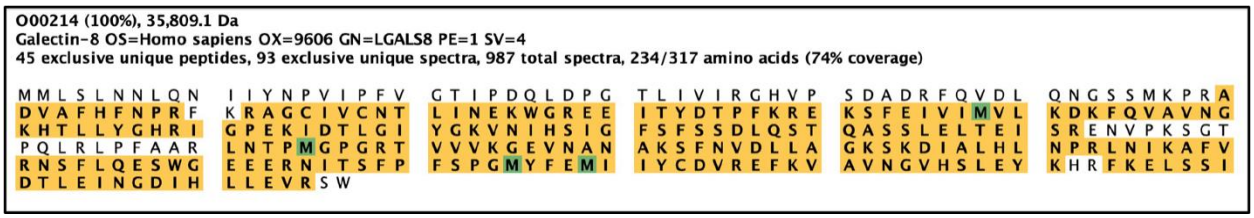


Fig. S8. Galectin-8 MS/MS results

Table S3. Galectin-3 CRD 0.1 mM IPTG Purification Summary

Step	Total Protein (mg)	Yield (%)	Purity (%)
Crude Lysate	182	100	2
Cleared Lysate	110	60	3
Ni-NTA Eluent (pooled)	3	2	95+

Cellular Concentration of Galectin-3 CRD from Leaky Expression

Final OD_{600} for expression cultures grown in LB media induced with 0.1 mM IPTG were typically 2.3. With an OD_{600} = 1.0 representing 8.0 x 10^8 *E. coli* cells/mL of culture [2], we arrive at a total cell count of 1.84 x 10^9 cells/mL in the IPTG-induced cultures. Given that *E. coli* cells have an average cell volume of 1 fL [3], we estimate the total cellular volume in 1 mL of culture at 1.84 x 10^{-3} mL (=1.84 x 10^9 cells)(10^{-15} L/cell)(10^3 ml/L). From our experiments, 1L of culture yields 7.4 mg of Galectin-3 CRD, which is 7.4 x 10^{-3} mg/mL. Since each mL of culture represents a cellular volume of 1.84 x 10^{-3} mL, the actual concentration of Galectin-3 CRD within the cell

is 4.02 mg/mL (=7.4 x 10^{-3} mg/1.84 x 10^{-3} ml). With a molar mass of ~16.5 kDa, the cellular concentration of Galectin-3 CRD in LB cultures induced with 0.1 mM IPTG is ~244 µM (= 4 g/L ÷ 16,500 g/mole). Relative to the 64 µM K_D, the cellular concentration of Gal-3 CRD from leaky expression is thus ~ 4 times greater than the dissociation constant for IPTG. This suggests leaky Gal-3 CRD is indeed present at concentrations high enough to sequester cytosolic IPTG and inhibit induction at lower IPTG concentrations.

Table S4. Galectin-3 CRD Autoinduction Purification Summary

Step	**Total Protein (mg)**	**Yield (%)**	**Purity (%)**
Crude Lysate	250	100	8
Cleared Lysate	160	64	13
Ni-NTA Eluent (pooled)	22	9	95+
Superdex 75 Column (pooled peak)	20	8	99+

Table S5. Mean yields of Gal-3 CRD with various induction protocols.

Induction Method	mg Gal-3 CRD/g cell pellet	mg Gal-3 CRD/L media
0 mM IPTG	1.4 ± 0.4[a]	6.8 ± 0.6
0.1 mM IPTG	1.5 ± 0.3	7.4 ± 0.6
0.5 mM IPTG	1.4 ± 0.2	7.0 ± 0.5
1.0 mM IPTG	1.4 ± 0.1	6.9± 0.5
2.0 mM IPTG	1.8 ± 0.1	9.0 ± 0.6
5.0 mM IPTG	2.6 ± 0.1	13.0 ± 0.8
Autoinduction	10.8 ± 1.0	102.7 ± 2.5

[a]Standard deviations are measured from at least 2 purifications.

References

1. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry*. Analytical Chemistry, 2003. **75**(17): p. 4646-4658.

2. Myers, J.A., B.S. Curtis, and W.R. Curtis, *Improving accuracy of cell and chromophore concentration measurements using optical density*. BMC Biophys, 2013. **6**(1): p. 4.

3. Kubitschek, H.E. and J.A. Friske, *Determination of bacterial cell volume with the Coulter Counter*. J Bacteriol, 1986. **168**(3): p. 1466-7.

CUMULATIVE REFERENCES CITED

1. Brussaard, C.P.D., A.-C. Baudoux, and F. Rodríguez-Valera, Marine Viruses, in The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential, L.J. Stal and M.S. Cretoiu, Editors. 2016, Springer International Publishing: Cham. p. 155-183.

2. Enikeeva, F.N., K.V. Severinov, and M.S. Gelfand, Restriction–modification systems and bacteriophage invasion: Who wins? Journal of Theoretical Biology, 2010. 266(4): p. 550-559.

3. Xu, T., et al., A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella. Nucleic Acids Research, 2010. 38(20): p. 7133-7141.

4. Labrie, S.J., J.E. Samson, and S. Moineau, Bacteriophage resistance mechanisms. Nature Reviews Microbiology, 2010. 8(5): p. 317-327.

5. Nordström, K. and A. Forsgren, Effect of protein A on adsorption of bacteriophages to Staphylococcus aureus. J Virol, 1974. 14(2): p. 198-202.

6. Hammad, A.M.M., Evaluation of alginate-encapsulated Azotobacter chroococcum as a phage-resistant and an effective inoculum. Journal of Basic Microbiology, 1998. 38(1): p. 9-16.

7. Destoumieux-Garzón, D., et al., The iron–siderophore transporter FhuA is the receptor for the antimicrobial peptide microcin J25: role of the microcin Val11–Pro16 ß-hairpin region in the recognition mechanism. Biochemical Journal, 2005. 389(3): p. 869-876.

8. Hille, F., et al., The Biology of CRISPR-Cas: Backward and Forward. Cell, 2018. 172(6): p. 1239-1259.

9. Sorek, R., C.M. Lawrence, and B. Wiedenheft, CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry, 2013. 82(1): p. 237-266.

10. Barrangou, R. and P. Horvath, A decade of discovery: CRISPR functions and applications. Nat Microbiol, 2017. 2: p. 17092.

11. Jackson, S.A., et al., CRISPR-Cas: Adapting to change. Science, 2017. 356(6333).

12. Koonin, E.V. and K.S. Makarova, Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci, 2019. 374(1772): p. 20180087.

13. 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.

14. McGinn, J. and L.A. Marraffini, Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat Rev Microbiol, 2019. 17(1): p. 7-12.

15. Mohanraju, P., et al., Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science, 2016. 353(6299): p. aad5147.

16. Barrangou, R., et al., CRISPR provides acquired resistance against viruses in prokaryotes. Science (New York, N.Y.), 2007. 315(5819): p. 1709-12.

17. Bolotin, A., et al., Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 2005. 151(Pt 8): p. 2551-61.

18. Groenen, P.M., et al., Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. Mol Microbiol, 1993. 10(5): p. 1057-65.

19. Ishino, Y., et al., Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol, 1987. 169(12): p. 5429-33.

20. Jansen, R., et al., Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol, 2002. 43(6): p. 1565-75.

21. Makarova, K.S., et al., A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology direct, 2006. 1: p. 7.

22. Marraffini, L.A. and E.J. Sontheimer, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science (New York, N.Y.), 2008. 322(5909): p. 1843-5.

23. Mojica, F.J., et al., Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol, 2005. 60(2): p. 174-82.

24. 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.

25. 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, 2005. 151(Pt 3): p. 653-63.

26. Sorek, R., C.M. Lawrence, and B. Wiedenheft, CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem, 2013. 82: p. 237-66.

27. Hille, F., et al., The Biology of CRISPR-Cas: Backward and Forward. Cell, 2018. 172(6): p. 1239-1259.

28. Charbonneau, A.A., C.C. Gauvin, and C.M. Lawrence, CRISPR-Cas, A Prokaryotic Adaptive Immune System, in Encyclopedia of Biological Chemistry, 3rd Edition, J. Joseph, Editor. 2021, Oxford: Elsevier. p. 717-741.

29. Anders, C., et al., Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature, 2014. 513(7519): p. 569-73.

30. Jiang, W., et al., Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. PLOS Genetics, 2013. 9(9): p. e1003844.

31. Stern, A., et al., Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics, 2010. 26(8): p. 335-340.

32. Agari, Y., et al., Transcription Profile of Thermus thermophilus CRISPR Systems after Phage Infection. Journal of Molecular Biology, 2010. 395(2): p. 270-281.

33. Fusco, S., et al., Transcriptome analysis of Sulfolobus solfataricus infected with two related fuselloviruses reveals novel insights into the regulation of CRISPR-Cas system. Biochimie, 2015. 118: p. 322-332.

34. Perez-Rodriguez, R., et al., Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. Molecular Microbiology, 2011. 79(3): p. 584-599.

35. Young, J.C., et al., Phage-Induced Expression of CRISPR-Associated Proteins Is Revealed by Shotgun Proteomics in Streptococcus thermophilus. PLOS ONE, 2012. 7(5): p. e38077.

36. Jore, M.M., et al., Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol, 2011. In Press.

37. 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). Journal of Biological Chemistry, 2011. 286(24): p. 21643-21656.

38. Sashital, D.G., M. Jinek, and J.A. Doudna, An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat Struct Mol Biol, 2011. 18(6): p. 680-7.

39. Charbonneau, A.A., C.C. Gauvin, and C.M. Lawrence, Methods | CRISPR-Cas, A Prokaryotic Adaptive Immune System, in Encyclopedia of Biological Chemistry III (Third Edition), J. Jez, Editor. 2021, Elsevier: Oxford. p. 717-741.

40. Brouns, S.J., et al., Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (New York, N.Y.), 2008. 321(5891): p. 960.

41. Deltcheva, E., et al., CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 2011. 471(7340): p. 602-7.

42. Hale, C.R., et al., RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell, 2009. 139(5): p. 945-56.

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

44. Sapranauskas, R., et al., The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res, 2011. 39(21): p. 9275-82.

45. Sakai, H.D. and N. Kurosawa, Saccharolobus caldissimus gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of Sulfolobus solfataricus as Saccharolobus solfataricus comb. nov. and Sulfolobus shibatae

as Saccharolobus shibatae comb. nov. Int J Syst Evol Microbiol, 2018. 68(4): p. 1271-1278.

46. Charbonneau, A.A., et al., Cyclic Tetra-Adenylate (cA4) Recognition by Csa3; Implications for an Integrated Class 1 CRISPR-Cas Immune Response in Saccharolobus solfataricus. Biomolecules, 2021. 11(12): p. 1852.

47. Kraft, P., et al., Structure of D-63 from Sulfolobus Spindle-Shaped Virus 1: Surface Properties of the Dimeric Four-Helix Bundle Suggest an Adaptor Protein Function. Journal of Virology, 2004. 78(14): p. 7438-7442.

48. Larson, E.T., et al., Structure of A197 from Sulfolobus Turreted Icosahedral Virus: a Crenarchaeal Viral Glycosyltransferase Exhibiting the GT-A Fold. Journal of Virology, 2006. 80(15): p. 7636-7644.

49. Lintner, N.G., et al., The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System. Journal of Molecular Biology, 2011. 405(4): p. 939-955.

50. 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.

51. Menon, S.K., et al., The crystal structure of D212 from sulfolobus spindle-shaped virus ragged hills reveals a new member of the PD-(D/E)XK nuclease superfamily. J Virol, 2010. 84(12): p. 5890-7.

52. Lintner, N.G., et al., The Structure of the CRISPR-Associated Protein Csa3 Provides Insight into the Regulation of the CRISPR/Cas System. J Mol Biol, 2011. 405(4): p. 939-955.

53. Lillestøl, R.K., et al., CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Molecular Microbiology, 2009. 72(1): p. 259-272.

54. Zink, I.A., E. Wimmer, and C. Schleper, Heavily Armed Ancestors: CRISPR Immunity and Applications in Archaea with a Comparative Analysis of CRISPR Types in Sulfolobales. Biomolecules, 2020. 10(11).

55. Hayes, R.P., et al., Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. Nature, 2016. 530(7591): p. 499-503.

56. Xiao, Y., et al., Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell, 2017. 170(1): p. 48-60 e11.

57. Jackson, R.N., et al., Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science, 2014. 345(6203): p. 1473-9.

58. Semenova, E., et al., Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A, 2011. 108(25): p. 10098-103.

59. Mulepati, S., A. Heroux, and S. Bailey, Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science, 2014. 345(6203): p. 1479-84.

60. Osawa, T., et al., Crystal Structure of the CRISPR-Cas RNA Silencing Cmr Complex Bound to a Target Analog. Mol Cell, 2015. 58(3): p. 418-30.

61. Staals, R.H., et al., RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol Cell, 2014. 56(4): p. 518-30.

62. Taylor, D.W., et al., Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science, 2015. 348(6234): p. 581-5.

63. Zhao, H., et al., Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature, 2014. 515(7525): p. 147-50.

64. Deng, L., et al., A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol, 2013. 87(5): p. 1088-99.

65. Elmore, J.R., et al., Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev, 2016. 30(4): p. 447-59.

66. Estrella, M.A., F.T. Kuo, and S. Bailey, RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev, 2016. 30(4): p. 460-70.

67. Kazlauskiene, M., et al., Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Mol Cell, 2016. 62(2): p. 295-306.

68. Samai, P., et al., Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell, 2015. 161(5): p. 1164-1174.

69. Tamulaitis, G., et al., Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell, 2014. 56(4): p. 506-17.

70. Haft, D., et al., A Guild of 45 CRISPR-Associated (Cas) Protien Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. Plos Computational Biology, 2005. 1(6): p. 474-483.

71. Makarova, K.S. and E.V. Koonin, Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol, 2015. 1311: p. 47-75.

72. Athukoralage, J.S., et al., Tetramerisation of the CRISPR ring nuclease Crn3/Csx3 facilitates cyclic oligoadenylate cleavage. Elife, 2020. 9.

73. Makarova, K.S., et al., CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Front Genet, 2014. 5: p. 102.

74. Makarova, K.S., et al., Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res, 2020. 48(16): p. 8828-8847.

75. Anantharaman, V., et al., Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biol Direct, 2013. 8: p. 15.

76. Kim, Y.K., Y.G. Kim, and B.H. Oh, Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of Pyrococcus furiosus. Proteins, 2013. 81(2): p. 261-70.

77. Sheppard, N.F., et al., The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. RNA, 2016. 22(2): p. 216-24.

78. Yan, X., W. Guo, and Y.A. Yuan, Crystal structures of CRISPR-associated Csx3 reveal a manganese-dependent deadenylation exoribonuclease. RNA Biol, 2015. 12(7): p. 749-60.

79. Topuzlu, E. and C.M. Lawrence, Recognition of a pseudo-symmetric RNA tetranucleotide by Csx3, a new member of the CRISPR associated Rossmann fold superfamily. RNA Biol, 2016. 13(2): p. 254-7.

80. Zhu, X. and K. Ye, Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. FEBS Lett, 2012. 586(6): p. 939-45.

81. Kazlauskiene, M., et al., A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science, 2017. 357(6351): p. 605-609.

82. Niewoehner, O., et al., Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature, 2017. 548(7669): p. 543-548.

83. Liu, T., et al., Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Res, 2015. 43(2): p. 1044-55.

84. He, F., et al., CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b. Nucleic Acids Res, 2017. 45(4): p. 1902-1913.

85. Liu, T., et al., Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res, 2017. 45(15): p. 8978-8992.

86. Madeira, F., et al., Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic acids research, 2022. 50(W1): p. W276-W279.

87. Peng, N., et al., A synthetic arabinose-inducible promoter confers high levels of recombinant protein expression in hyperthermophilic archaeon Sulfolobus islandicus. Appl Environ Microbiol, 2012. 78(16): p. 5630-7.

88. Dlakić, M., HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch. Bioinformatics, 2009. 25(23): p. 3071-6.

89. Studier, F.W., Protein production by auto-induction in high density shaking cultures. Protein expression and purification, 2005. 41(1): p. 207.

90. Bradford, M.M., A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical Biochemistry, 1976. 72(1-2): p. 248-254.

91. Brown, S., et al., Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR-Cas that degrades the second messenger cA4. J Biol Chem, 2020. 295(44): p. 14963-14972.

92. Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna, Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. RNA (New York, N.Y.), 2012. 18(4): p. 661-672.

93. Otwinowski, Z. and W. Minor, [20] Processing of X-ray diffraction data collected in oscillation mode, in Methods in Enzymology. 1997, Elsevier. p. 307-326.

94. Mccoy, A.J., et al., Phasercrystallographic software. Journal of Applied Crystallography, 2007. 40(4): p. 658-674.

95. Adams, P.D., et al., PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D Biological Crystallography, 2010. 66(2): p. 213-221.

96. Emsley, P., et al., Features and development ofCoot. Acta Crystallographica Section D Biological Crystallography, 2010. 66(4): p. 486-501.

97. Painter, J. and E.A. Merritt, Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr D Biol Crystallogr, 2006. 62(Pt 4): p. 439-50.

98. Chen, V.B., et al., MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D Biological Crystallography, 2010. 66(1): p. 12-21.

99. Delano, W.L., The PyMOL Molecular Graphics System. http://www.pymol.org, 2002.

100. Athukoralage, J.S., et al., Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. Nature, 2018. 562(7726): p. 277-280.

101. Athukoralage, J.S., et al., An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. Nature, 2020. 577(7791): p. 572-575.

102. Athukoralage, J.S., et al., The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling. Elife, 2020. 9.

103. Athukoralage, J.S., et al., A Type III CRISPR Ancillary Ribonuclease Degrades Its Cyclic Oligoadenylate Activator. J Mol Biol, 2019. 431(15): p. 2894-2899.

104. Garcia-Doval, C., et al., Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6. Nat Commun, 2020. 11(1): p. 1596.

105. Jia, N., et al., CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity. Mol Cell, 2019. 75(5): p. 944-956 e6.

106. Richardson, J.S., et al., RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). RNA, 2008. 14(3): p. 465-81.

107. Hong, M., et al., Structure of an OhrR-ohrA operator complex reveals the DNA binding mechanism of the MarR family. Mol Cell, 2005. 20(1): p. 131-41.

108. Peng, W., et al., An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. Nucleic Acids Res, 2015. 43(1): p. 406-17.

109. Baker-Austin, C. and M. Dopson, Life in acid: pH homeostasis in acidophiles. Trends Microbiol, 2007. 15(4): p. 165-71.

110. Moll, R. and G. Schäfer, Chemiosmotic H+ cycling across the plasma membrane of the thermoacidophilic archaebacterium Sulfolobus acidocaldarius. FEBS Lett, 1988. 232(2): p. 359-363.

111. Rostol, J.T., et al., The Card1 nuclease provides defence during type III CRISPR immunity. Nature, 2021. 590(7847): p. 624-629.

112. Ye, Q., et al., CRISPR-Associated Factor Csa3b Regulates CRISPR Adaptation and Cmr-Mediated RNA Interference in Sulfolobus islandicus. Front Microbiol, 2020. 11: p. 2038.

113. Rouillon, C., et al., Control of cyclic oligoadenylate synthesis in a type III CRISPR system. Elife, 2018. 7.

114. Liu, Z., et al., A CRISPR-associated factor Csa3a regulates DNA damage repair in Crenarchaeon Sulfolobus islandicus. Nucleic Acids Res, 2020. 48(17): p. 9681-9693.

115. Cohen, D., et al., Cyclic GMP-AMP signalling protects bacteria against viral infection. Nature, 2019. 574(7780): p. 691-695.

116. Lowey, B., et al., CBASS Immunity Uses CARF-Related Effectors to Sense 3'-5'- and 2'-5'-Linked Cyclic Oligonucleotide Signals and Protect Bacteria from Phage Infection. Cell, 2020. 182(1): p. 38-49 e17.

117. Lau, R.K., et al., Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity. Mol Cell, 2020. 77(4): p. 723-733 e6.

118. Burroughs, A.M., et al., Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. Nucleic Acids Res, 2015. 43(22): p. 10633-54.

119. Whiteley, A.T., et al., Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. Nature, 2019. 567(7747): p. 194-199.

120. Shmakov, S.A., et al., Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc Natl Acad Sci U S A, 2018. 115(23): p. E5307-E5316.

121. White, M.F., Bacteria SAVED from Viruses. Cell, 2020. 182(1): p. 5-6.

122. Marino, K.V., et al., Targeting galectin-driven regulatory circuits in cancer and fibrosis. Nature Reviews Drug Discovery, 2023. 22(4): p. 295-316.

123. Navarro, P., et al., Impact of Galectins in Resistance to Anticancer Therapies. Clinical cancer research : an official journal of the American Association for Cancer Research, 2020.

124. Arthur, C.M., et al., Evolving Mechanistic Insights into Galectin Functions, in Galectins: Methods and Protocols, S.R. Stowell and R.D. Cummings, Editors. 2015. p. 1-35.

125. Troncoso, M.F., et al., The Universe of Galectin-Binding Partners and Their Functions in Health and Disease. The Journal of biological chemistry, 2023: p. 105400-105400.

126. Caballero, G.G., et al., How galectins have become multifunctional proteins. Histology and Histopathology, 2020. 35(6): p. 509-539.

127. Sciacchitano, S., et al., Galectin-3: One Molecule for an Alphabet of Diseases, from A to Z. International Journal of Molecular Sciences, 2018. 19(2): p. 59.

128. Nio-Kobayashi, J., Histological Mapping and Subtype-Specific Functions of Galectins in Health and Disease. Trends in Glycoscience and Glycotechnology, 2018. 30(172): p. SE89-SE96.

129. Dings, R.P.M., et al., Simulating cellular galectin networks by mixing galectins in vitro reveals synergistic activity. Biochemistry and biophysics reports, 2021. 28: p. 101116-101116.

130. Stowell, S.R. and R.D. Cummings, Galectins: Methods and Protocols. 2015, New York: Springer Science.

131. Cummings, R.D. and F.T. Liu, Galectins, in Essentials of Glycobiology. 2nd ed., A. Varki, et al., Editors. 2009, Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY).

132. Gabius, H.J., et al., What is the Sugar Code? Chembiochem.

133. Di Lella, S., et al., When Galectins Recognize Glycans: From Biochemistry to Physiology and Back Again. Biochemistry, 2011. 50(37): p. 7842-7857.

134. Bernhard, S.P., et al., Protein aggregation nucleated by functionalized dendritic polyglycerols. Polymer Chemistry, 2020. 11(23): p. 3849-3862.

135. van Hattum, H., et al., Tuning the Preference of Thiodigalactoside- and Lactosamine-Based Ligands to Galectin-3 over Galectin-1. Journal of Medicinal Chemistry, 2013. 56(3): p. 1350-1354.

136. van Scherpenzee, M., et al., Synthesis and Evaluation of New Thiodigalactoside-Based Chemical Probes to Label Galectin-3. Chembiochem, 2009. 10(10): p. 1724-1733.

137. Giguère, D., et al., Aryl O- and S-galactosides and lactosides as specific inhibitors of human galectins-1 and -3:: Role of electrostatic potential at O-3. Bioorganic & Medicinal Chemistry Letters, 2006. 16(6): p. 1668-1672.

138. Cumpstey, I., et al., Synthesis of a phenyl thio-ß-D-galactopyranoside library from 1,5-difluoro-2,4-dinitrobenzene:: discovery of effcient and selective monosaccharide inhibitors of galectin-7. Organic & Biomolecular Chemistry, 2005. 3(10): p. 1922-1932.

139. Dam, T.K., et al., Galectins bind to the multivalent glycoprotein asialofetuin with enhanced affinities and a gradient of decreasing binding constants. Biochemistry, 2005. 44(37): p. 12564-12571.

140. Poland, P.A., C.L. Kinlough, and R.P. Hughey, Cloning, Expression, and Purification of Galectins for In Vitro Studies. Galectins: Methods and Protocols, 2015. 1207: p. 37-49.

141. Poland, P.A., C.L. Kinlough, and R.P. Hughey, Cloning, Expression, and Purification of Galectins for In Vitro Studies. Methods in molecular biology (Clifton, N.J.), 2022. 2442: p. 41-54.

142. Prato, C.A., et al., Purification of Recombinant Galectins Expressed in Bacteria. Star Protocols, 2020. 1(3).

143. Dey, C., P. Palm, and L. Elling, Characterization of Galectin Fusion Proteins with Glycoprotein Affinity Columns and Binding Assays. Molecules, 2023. 28(3).

144. Bohari, M.H., et al., Structure-Based Design of a Monosaccharide Ligand Targeting Galectin-8. Chemmedchem, 2018. 13(16): p. 1664-1672.

145. Chien, C.T.H., et al., Lactose Binding Induces Opposing Dynamics Changes in Human Galectins Revealed by NMR-Based Hydrogen-Deuterium Exchange. Molecules, 2017. 22(8).

146. Zhang, H., et al., Discovery of N-Arylsulfonyl-Indole-2-Carboxamide Derivatives as Galectin-3 and Galectin-8 C-Terminal Domain Inhibitors. Acs Medicinal Chemistry Letters, 2023. 14(9): p. 1257-1265.

147. Yoshida, H., et al., X-ray structure of a protease-resistant mutant form of human galectin-8 with two carbohydrate recognition domains. Febs Journal, 2012. 279(20): p. 3937-3951.

148. Si, Y., et al., Linker remodels human Galectin-8 structure and regulates its hemagglutination and pro-apoptotic activity. International Journal of Biological Macromolecules, 2023. 245.

149. Bernhard, S.P., et al., Time-Dependent Fluorescence Spectroscopy to Quantify Complex Binding Interactions. ACS Omega, 2020. 5(45): p. 29017-29024.

150. Studier, F.W., Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification, 2005. 41(1): p. 207-234.

151. Menon, S.K., et al., The Crystal Structure of D212 from Sulfolobus Spindle-Shaped Virus Ragged Hills Reveals a New Member of the PD-(D/E)XK Nuclease Superfamily. Journal of Virology, 2010. 84(12): p. 5890-5897.

152. Lintner, N.G., et al., The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. J Mol Biol, 2011. 405(4): p. 939-55.

153. Bertani, G., STUDIES ON LYSOGENESIS I. Journal of Bacteriology, 1951. 62(3): p. 293-300.

154. Brown, S., et al., Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR-Cas that degrades the second messenger cA(4). J Biol Chem, 2020. 295(44): p. 14963-14972.

155. Petersen, K.J., et al., Fluorescence lifetime plate reader: Resolution and precision meet high-throughput. Review of Scientific Instruments, 2014. 85(11): p. 113101.

156. Muretta, J.M., et al., High-performance time-resolved fluorescence by direct waveform recording. Review of Scientific Instruments, 2010. 81(10): p. 103101.

157. Schlick, K.H., et al., Characterization of Protein Aggregation via Intrinsic Fluorescence Lifetime. Journal of the American Chemical Society, 2009. 131(46): p. 16608-16609.

158. Ennist, J.H., et al., Chemoenzymatic Synthesis of Galectin Binding Glycopolymers. Bioconjugate Chemistry, 2018. 29(12): p. 4030-4039.

159. Goodman, C.K., et al., Multivalent scaffolds induce galectin-3 aggregation into nanoparticles. Beilstein Journal of Organic Chemistry, 2014. 10: p. 1570-1577.

160. Kubitschek, H.E. and J.A. Friske, DETERMINATION OF BACTERIAL-CELL VOLUME WITH THE COULTER-COUNTER. Journal of Bacteriology, 1986. 168(3): p. 1466-1467.

161. Myers, J.A., B.S. Curtis, and W.R. Curtis, Improving accuracy of cell and chromophore concentration measurements using optical density. Bmc Biophysics, 2013. 6.

162. Brown, S., et al., Csx3 is a cyclic oligonucleotide phosphodiesterase associated with type III CRISPR–Cas that degrades the second messenger cA4. Journal of Biological Chemistry, 2020. 295(44): p. 14963-14972.

163. Monod, J., The growth of bacterial cultures. Annual review of microbiology, 1949. 3(1): p. 371-349.

164. Stulke, J. and W. Hillen, Carbon catabolite repression in bacteria. Curr Opin Microbiol, 1999. 2(2): p. 195-201.

165. Bettenbrock, K., et al., A quantitative approach to catabolite repression in Escherichia coli. J Biol Chem, 2006. 281(5): p. 2578-84.

166. Studier, F.W., Protein production by auto-induction in high density shaking cultures. Protein Expr Purif, 2005. 41(1): p. 207-34.

167. Voet, D. and J.G. Voet, Biochemistry. 4th ed. 2011, Hoboken, NJ: John Wiley & Sons. xxv, 1428, 53 p.

168. Studier, F.W. and B.A. Moffatt, Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol, 1986. 189(1): p. 113-30.

169. Studier, F.W., et al., Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol, 1990. 185: p. 60-89.

170. Auchincloss, L.C., et al., Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. Cbe-Life Sciences Education, 2014. 13(1): p. 29-40.

171. Buchanan, A.J. and G.R. Fisher, Current Status and Implementation of Science Practices in Course-Based Undergraduate Research Experiences (CUREs): A Systematic Literature Review. Cbe-Life Sciences Education, 2022. 21(4).

172. Watts, F.M. and J.-M.G. Rodriguez, A Review of Course-Based Undergraduate Research Experiences in Chemistry. Journal of Chemical Education, 2023. 100(9): p. 3261-3275.

173. Bangera, G. and S.E. Brownell, Course-Based Undergraduate Research Experiences Can Make Scientific Research More Inclusive. Cbe-Life Sciences Education, 2014. 13(4): p. 602-606.

174. Rodenbusch, S.E., et al., Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees. Cbe-Life Sciences Education, 2016. 15(2).

175. Estrada, M., et al., Improving Underrepresented Minority Student Persistence in STEM. Cbe-Life Sciences Education, 2016. 15(3).

176. Groover, S., J. Legleiter, and E.E. Battin, Huntingtin Protein Purification and Experimentation: An Innovative Undergraduate Research-Based Biochemistry Experiment. Journal of Chemical Education, 2021. 98(9): p. 3011-3018.

177. Makarova, K.S., et al., Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol, 2011. 9(6): p. 467-77.

178. Makarova, K.S., et al., An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol, 2015. 13(11): p. 722-36.

179. Liu, T., et al., Coupling transcriptional activation of CRISPR–Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Research, 2017. 45(15): p. 8978-8992.

180. Ye, Q., et al., CRISPR-Associated Factor Csa3b Regulates CRISPR Adaptation and Cmr-Mediated RNA Interference in Sulfolobus islandicus. Frontiers in Microbiology, 2020. 11.

181. Nesvizhskii, A.I., et al., A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry, 2003. 75(17): p. 4646-4658.

182. Myers, J.A., B.S. Curtis, and W.R. Curtis, Improving accuracy of cell and chromophore concentration measurements using optical density. BMC Biophys, 2013. 6(1): p. 4.

183. Kubitschek, H.E. and J.A. Friske, Determination of bacterial cell volume with the Coulter Counter. J Bacteriol, 1986. 168(3): p. 1466-7.

184. Makarova, K.S., et al., Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biol Direct, 2012. 7: p. 40.

185. 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.

186. Shah, S.A., et al., Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biol, 2019. 16(4): p. 530-542.

187. Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997. 25(17): p. 3389-402.

188. Madeira, F., et al., The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res, 2019. 47(W1): p. W636-W641.

189. Krissinel, E. and K. Henrick, Inference of macromolecular assemblies from crystalline state. J Mol Biol, 2007. 372(3): p. 774-97.

190. Bennett, M.J., M.P. Schlunegger, and D. Eisenberg, 3D domain swapping: a mechanism for oligomer assembly. Protein Sci, 1995. 4(12): p. 2455-68.

191. Okonechnikov, K., et al., Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics, 2012. 28(8): p. 1166-7.

192. Zimmermann, L., et al., A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol, 2018. 430(15): p. 2237-2243.

193. Bradford, M.M., A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem., 1976. 72: p. 248-54.

194. Petersen, K.J., et al., Fluorescence lifetime plate reader: resolution and precision meet high-throughput. Rev Sci Instrum, 2014. 85(11): p. 113101.

195. Muretta, J.M., et al., High-performance time-resolved fluorescence by direct waveform recording. Rev Sci Instrum, 2010. 81(10): p. 103101.

196. Schlick, K.H., et al., Characterization of protein aggregation via intrinsic fluorescence lifetime. J Am Chem Soc, 2009. 131(46): p. 16608-9.

197. Otwinowski, Z. and W. Minor, [20] Processing of X-ray diffraction data collected in oscillation mode, in Methods in Enzymology. 1997, Academic Press. p. 307-326.

198. McCoy, A.J., et al., Phaser crystallographic software. Journal of Applied Crystallography, 2007. 40: p. 658-674.

199. Adams, P.D., et al., PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr, 2010. D66: p. 213-221.

200. Emsley, P., et al., Features and development of Coot. Acta Crystallographica Section D-Biological Crystallography, 2010. 66: p. 486-501.

201. Adams, P.D., et al., PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D-Structural Biology, 2010. 66: p. 213-221.

202. Chen, V.B., et al., MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D-Structural Biology, 2010. 66: p. 12-21.

203. DeLano, W.L. The PyMOL Molecular Graphics System. http://www.pymol.org. 2002; Available from: http://www.pymol.org.

204. Sternberg, S.H., R.E. Haurwitz, and J.A. Doudna, Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. RNA, 2012. 18(4): p. 661-72.