INTERSECTION OF SARS-CoV-2 AND CRISPR-CAS DEFENSE SYSTEMS

by

Tanner Roy Wiegand

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

of

Doctor of Philosophy

in

Microbiology and Cell Biology

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2021

©COPYRIGHT

by

Tanner Roy Wiegand

2021

All Rights Reserved

ACKNOWLEDGEMENTS

Much like investigating any complex scientific question, completing a doctoral degree requires one to lean on and learn from a talented and tenacious team. I have been fortunate in this regard and would first like to thank Blake Wiedenheft for his mentorship and encouragement in his role as my advisor. Blake likes to joke that I arrived at his lab not knowing which end of the pipette to hold. As I reflect on all I've learned from Blake and my other mentors, I appreciate more and more the kernel of truth embedded in this jest. Blake's uncompromising curiosity challenges his students to understand the mechanics of the biological world around them. Working with him has made me a better writer, presenter, and scientist.

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

Finally, I want to acknowledge those whose love and support were a haven when experiments failed, or nature refused to reveal her secrets. My friends keep me sane and grounded, and they are bright lights in my life. A special thanks to my parents, Maury and Bruce, and my siblings, Jackie, Brady, Kenenni, and Hanni. Family has always been a cornerstone in my life, and I couldn't have completed this work without you.

TABLE OF CONTENTS

1. REPRODUCIBLE ANTIGEN RECOGNITION BY THE TYPE I-F CRISPR SYSTEM 1

Contribution of Authors and Co-Authors 1
Manuscript information 2
Abstract 4
Introduction 5
Materials and Methods 6
Adaptation Assays in *Escherichia coli* 6
High-throughput Sequencing 7
Spacer Sequence Analysis and Statistics 8
Results 9
Csy-PAM interactions Increase I-F CRISPR Adaptation Efficiency 9
Cas1-2/3 are Necessary and Sufficient for Inefficient I-F Spacer Acquisition 9
The Csy Complex Increases the Percentage of PAM-proximal Spacers 11
Spacers are Preferentially Acquired from the Terminus of Plasmid Replication 14
Some Spacers are Reproducibly Acquired 17
Discussion 18
Acknowledgments 22
References 23
Supplemental Figures 27
Supplemental Tables 28

2. DISTRIBUTION AND PHASING OF SEQUENCES MOTIFS THAT FACILITATE CRISPR ADAPTATION 30

Contribution of Authors and Co-Authors 30
Manuscript information 31
Abstract 33
Introduction 34
Results 37
Identification and Distribution of IHF and Upstream Motifs 37
IHF-binding Sites and Upstream Motif are Critical for Efficient Integration 42
Tn7-associated I-F3 leaders Contain Motifs needed for Efficient Integration 45
Phased Leader Motifs Facilitate CRISPR Adaptation 46
Discussion 48

TABLE OF CONTENTS CONTINUED

Acknowledgments 52
References 53
Supplemental Figures 61
Materials and Methods 68
Bacterial Strains 68
Modeling the I-F CRISPR Integration Intermediate Complex 68
Building a Database of CRISPR Leaders 69
Identification of Conserved DNA Motifs in CRISPR leaders 69
Phylogenetic Analyses 71
Plasmid Construction 71
IHF Expression and Purification 72
Cas1-2/3 Expression and Purification 73
Nucleic Acid Preparation 74
In vitro Integration Assays 75
High-throughput Sequencing of *in vitro* Integration Products 76
Quantification and Statistical Analysis 77

3. STRUCTURES AND STRATEGIES OF ANTI-CRISPR-MEDIATED IMMUNE SUPPRESSION 78

Contribution of Authors and Co-Authors 78
Manuscript information 79
Abstract 81
Introduction 82
Timing is Everything 83
Stoichiometric Inhibitors of CRISPR Defense 87
Anti-CRISPR Proteins that Masquerade as dsDNA 87
Acr-Mediated Dimerization 89
Why Two Acrs? 93
Enzymatic Suppressors of CRISPR Immunity 94
Covalent Modification of the Surveillance Complex 94
Degradation of Second Messenger Molecules 96
The Future of Anti-CRISPR 97
Acknowledgments 99
References 100

4. CRISPR SURVEILLANCE TURNS TRANSPOSON TAXI 110

Contribution of Authors and Co-Authors 110
Manuscript information 111
Four Reports Capture Structures of TniQ in Association with a crRNA-guided

TABLE OF CONTENTS CONTINUED

Surveillance Complex 113
Hell Breaks Loose 115
Acknowledgments 116
References 117

5. CAS9 SLIDE AND SEEK FOR PHAGE DEFENSE 119

Contribution of Authors and Co-Authors 119
Manuscript information 120
Abstract 122
Main Text 123
References 127

6. TEMPORAL DETECTION AND PHYLOGENETIC ASSESSMENT OF SARS-COV-2 IN MUNICIPAL WASTEWATER 128

Contribution of Authors and Co-Authors 128
Manuscript information 130
Abstract 132
Introduction 133
Results 133
Discussion 139
Limitations of Study 140
Acknowledgments 141
References 142
Supplemental Figures 147
Supplemental Tables 149
Materials and Methods 151
Wastewater Samples 151
Symptom Onset Data and Clinical Test Results 151
Wastewater sample processing and RNA extraction 152
Reverse Transcription Quantitative PCR (RT-qPCR) 152
RT-PCR and SARS-CoV-2 Genome Sequencing 153
SARS-CoV-2 Genome Assembly 154
Phylogenetic Analysis 155
Position-specific Mutation Analysis 155
Quantification and Statistical Analysis 155

7. SARS-COV-2 GENOMIC SURVEILLANCE IDENTIFIES NATURALLY OCCURING TRUNCATION OF ORF7A THAT LIMITS IMMUNE SUPPRESSION 157

TABLE OF CONTENTS CONTINUED

Contribution of Authors and Co-Authors 157
Manuscript information 159
Abstract 161
Introduction 162
Results 163
SARS-CoV-2 Genomic Surveillance Identifies Truncations of ORF7a 163
ORF7a Truncation has a Loss-of-Function Effect 166
ORF7a Mutation has no Collateral Effect on ORF7b 168
ORF7a$^{\Delta 115}$ Isolate Displays a Replication Defect 169
ORF7a$^{\Delta 115}$ Virus Induces Elevated IFN-I Response *in vitro* 171
Discussion 173
Limitations of Study 175
Acknowledgments 176
References 177
Supplemental Figures 184
Materials and Methods 188
Cell Cultures 188
Plasmids 188
Antibodies 188
Human Clinical Sample Collection and Preparation 189
Production and Titration of Coronavirus Stocks 189
Symptom Onset Data and Clinical Test Results 190
Quantitative Reverse Transcription PCR (qRT-PCR) 190
RT-PCR and SAR-CoV-2 Genome Sequencing 191
SARS-CoV-2 Genome Assembly 192
RT-PCR and Sanger Sequencing 192
Phylogenetic and ORF7a Mutational Analysis 193
Western Blot 194
Immunocytochemistry 195
IFN Inhibition Assay 195
SARS-CoV-2 Replication Assays 196
Type I Interferon Response Assay 197
Quantification and Statistical Analysis 198

8. INTRINSIC SIGNAL AMPLIFICATION BY TYPE III CRISPR-CAS SYSTEM PROVIDES A SEQUENCE-SPECIFIC SARS-COV-2 DIAGNOSTIC 199

Contribution of Authors and Co-Authors 199
Manuscript information 201
Abstract 203

TABLE OF CONTENTS CONTINUED

Introduction 204
Results 206
Sequence-specific Activation of Cas10 Polymerase Yields Three Detectable Products 206
Csm-based Direct Detection of SARS-CoV-2 RNA in Patient Samples 209
Testing Clinical Samples for SARS-CoV-2 using RT-LAMP and T7-Csm 211
Discussion 214
Limitations of Study 216
Acknowledgments 216
References 218
Supplemental Figures 224
Supplemental Tables 229
Materials and Methods 231
Bacterial Strains 231
Nucleic Acid Preparation 231
Plasmids 231
Protein Purifications 232
Type III CRISPR-based RNA Detection 234
Colorimetric CRISPR-Csm based Detection 235
Visible fluorometric CRISPR-Csm based Detection 236
RT-LAMP-T7-Csm 237
Human Clinical Sample Collection and Preparation 238
RT-qPCR 238
Bioinformatic Design of TtCsm crRNA Guides Targeting SARS-CoV-2 238
Sequencing of SARS-CoV-2 RNA Isolated from Patient Samples 239
Quantification and Statistical Analysis 240

9. FUNCTIONAL AND PHYLOGENETIC DIVERSITY OF CAS10 PROTEINS 241

Contribution of Authors and Co-Authors 241
Manuscript information 242
Abstract 244
Introduction 245
Materials and Methods 246
Identification of Cas10 proteins 246
Phylogenetic and Domain Composition Analyses 247
Determining the Presence of HD and Palm2 Catalytic Motifs 248

TABLE OF CONTENTS CONTINUED

Cas10 Protein Purification....248
Cas10 Cyclase Activity Assays....249
Cas10 Polymerase Activity Assays....250
Results....251
Phylogenetic Diversity and Proposed Evolutionary History of Cas10 Proteins....251
Difference in Cas10 HD and Palm Active Sites....253
Variations of Cas10 Domain Architecture....255
Cas10 cOA Production Requires All Components of the Type III Surveillance Complex....257
Diverse Cas10 proteins function as DNA-dependent DNA polymerases with varying efficiencies....259
Discussion....261
Acknowledgments....264
References....265
Supplemental Figures....270
Supplemental Tables....273

10. PHYLOGENETIC ANALYSIS OF CAS7 REVEALS GENETIC FUSIONS TOXIN-ANTITOXIN AND INNATE IMMUNITY COMPONENTS....274

Contribution of Authors and Co-Authors....274
Manuscript information....275
Abstract....277
Introduction....278
Materials and Methods....279
Cas7 Identification....279
Phylogenetic and Domain Architecture Analysis....279
Results....280
Cas7 Phylogeny Supports the Proposed Recent Origin of Type I CRISPR Systems....280
Cas7 Proteins are Genetically Fused to other Cas and CRISPR-associated Transposition Proteins....282
Fusions of Cas7 to Components of a HEPN-based Toxin-Antitoxin System....284
Cas7 proteins are Fused to Components of Innate Immune Systems....287
Discussion....290
Acknowledgments....292
References....294
Supplemental Tables....301

TABLE OF CONTENTS CONTINUED

11. GENERAL CONCLUSION AND OUTLOOK 302

Adaptation in CRISPR-Cas systems 303
SARS-CoV-2 and the COVID-19 Pandemic 304
Concluding Remarks 306
References 307

REFERENCES CITED 309

LIST OF TABLES

Table Page

S1.1. Primers used in these experiments 28

S1.2. Summary of High-throughput sequencing data 29

2.1. Subtype distribution of microbial CRISPRs and identified IHF-binding sites 38

2.2. Cas1 residues that interact with canonical I-E UM are only conserved in Cas1s associated with IHF-regulated I-E CRISPRs 42

S6.1. Total wastewater flow volumes and SARS-CoV-2 RNA concentration in composite samples 149

S6.2. Primer design used for RT-PCR 150

8.1. Comparison of clinical parameters of RT-LAMP-T7-Csm and FDA EUA-approved protocols of DETECTR and SHERLOCK diagnostics 214

S8.1. TtCsm guide RNAs 229

S8.2. Fluorescent reporter RNAs 229

S8.3. Primers to generate IVT RNA templates 230

S8.4. Primers for RT-LAMP 230

S8.5. Purified genomic nucleic acids 230

S9.1. Quantification of HD and Palm2 catalytic site motifs for different Cas10 subtypes 214

S10.1. Profiles HMMs from CasFinder used to identify Cas7 proteins in genomic and metagenomic databases 214

LIST OF FIGURES

Figure Page

1.1. Cas1–2/3 proteins are necessary and sufficient for new spacer acquisition, but Csy accelerates adaptation 10

1.2. Cas1-2/3 acquire mostly PAM-flanked spacers, but Csy increases the fidelity of PAM selection 13

1.3. Pearson correlation analysis of uniquely aligned pCas1-2/3 protospacer quantities and locations 15

S1.3. Protospacer hotspots are reproducible across multiple biological replicates 27

2.1. Spacing of sequence motifs in CRISPR leaders has mechanistic implications for the mechanism of spacer integration 36

2.2. IHF-directed CRISPR adaptation is widespread 41

2.3. IHF-binding sites and UM are critical for efficient integration 43

2.4. Phase-dependent CRISPR adaptation 47

S2.1. Naturally occurring indels preserve the phase of leader motifs 61

S2.2. DNA-interacting residues are conserved between *E. coli* and *P. aeruginosa* IHF proteins 62

S2.3. High-throughput sequencing and quantification of prespacer integration events 63

S2.4. IHF binds I-F leaders to recruit an upstream motif needed for efficient prespacer integration 65

S2.5. Summary of differential evolution rates between co-existing CRISPR loci 66

S2.6. Leader motifs involved in CRISPR evolution are phase sensitive 67

3.1. CRISPR-Cas adaptive immunity and structurally determined anti-CRISPRs (Acrs) 84

LIST OF FIGURES CONTINUED

Figure Page

3.2. Anti-CRISPR (Acr) and anti-CRISPR-associated (Aca) proteins 86

3.3. Anti-CRISPRs more closely mimic bound double-stranded DNA (dsDNA) than idealized B-form DNA 88

3.4. Acrs that oligomerize Cas protein targets 90

3.5. Enzymatic Acrs that inhibit type V-A CRISPR-Cas systems 95

4.1. Comparison of the I-F CRISPR defense system to the I-F Tn7-like system 114

5.1. PAM-mediated dwell times of Cas9 124

6.1. Detection and Quantification of SARS-CoV-2 in Wastewater and in the Community 135

6.2. Phylogenetic Analysis of SARS-CoV-2 Sequence Isolated from Wastewater 138

S6.1. Standard curves for absolute quantification of SARS-CoV-2 titers 147

S6.2. Sequencing and phylogenetics of SARS-CoV-2 in wastewater 148

7.1. SARS-CoV-2 genomic surveillance identifies global reoccurrence of ORF7a truncations 164

7.2. ORF7a truncation results in loss of function 167

7.3. The ORF7a$^{\Delta 115}$ mutation results in replication and immune suppression defects 170

S7.1. Phylogenetic and mutational analysis of ORF7a gene sequences 184

S7.2. Mapping TRS-B sites required for ORF7a and ORF7b discontinuous sgRNA transcription 185

S7.3. IFN-I response and gene set enrichment analysis 186

8.1. Detection of SARS-CoV-2 using the type III CRISPR-Cas system 207

LIST OF FIGURES CONTINUED

Figure Page

8.2. Screening and multiplexing crRNA guides 210

8.3. CRISPR-Csm-based detection of SARS-CoV-2 in clinical nasopharyngeal swab samples 212

S8.1. Purification of TtCsm complexes and TtCsm6 nuclease 224

S8.2. Csm-mediated synthesis of cA_4 is necessary for activation of the non-sequence specific nuclease Csm6 225

S8.3. Colorimetric detection of a specific RNA sequence 226

S8.4. Visible fluorometric detection of a specific RNA sequence 226

S8.5. Rapid and specific detection of SARS-CoV-2 by RT-LAMP-T7-Csm 227

S8.6. Standard curves for absolute quantification of SARS-CoV-2 titers 228

9.1. Cas10 proteins have variable domain compositions, with subtype-specific patterns 252

9.2. Cas10 alone is incapable of cOA-production but has DNA polymerase activity 258

S9.1. Fluorescence traces of Calcein assays to test cyclase activity with rNTPs 270

S9.2. Fluorescence traces of Calcein assays to test PfuCmr2 cyclase activity with dNTPs 271

S9.3. Fluorescence traces of Calcein assays to test cyclase activity with rATP only 270

10.1. Phylogenetic analysis of Cas7 sequences reveals novel genetic fusions 281

10.2. Genetic fusion of Cas7 to an antitoxin suggests repurposing of TA system components for CRISPR immunity 285

10.3. Cmr6-TIR-SAVED represents fusion of bacterial innate and adaptive immune components 289

ABSTRACT

Viral predators exploit cellular resources in all domains of life. To defend against these genetic invaders, bacteria and archaea have evolved adaptive immune systems comprised of clustered regularly interspaced short palindromic repeats (CRISPR) and their associated Cas proteins. In this dissertation, I investigate the biological mechanisms and biotechnological applications of CRISPR-Cas systems.

The sequences that interspace the eponymous repeats of CRISPR loci are derived from mobile genetic elements, including bacteriophages (i.e., viruses that infect bacteria). When the locus is transcribed into CRISPR-RNA, these spacer sequences guide nucleases to RNA or DNA molecules with complementary sequences, resulting in degradation of the target nucleic acid. While recent work has illuminated many details of CRISPR-RNA-guided surveillance and target interference, the process of new sequence adaptation remains more mysterious.

Initially, the goal of this research was to understand how new spacer sequences are acquired and integrated at CRISPR loci. High throughput sequencing of spacers acquired in *in vivo* adaptation assays revealed that some spacer sequences are reproducibly acquired in the I-F CRISPR system of *Pseudomonas aeruginosa*, and that the I-F CRISPR-guided surveillance complex enhances the efficiency of new spacer acquisition. We then used bioinformatic and *in vitro* acquisition assays to show that adaptation in many systems is dependent on the presence and phasing of sequence motifs in the transcriptional leaders of CRISPR loci. Collectively, these results expand our understanding of how CRISPR-Cas systems adapt to new threats.

Following the emergence of SARS-CoV-2, and the ensuing international COVID-19 pandemic, my research goals pivoted to developing methods to track the spread of this coronavirus and to understanding how it was evolving. Long read genomic sequencing was used to determine the likely evolutionary origin of SARS-CoV-2 samples isolated from wastewater and human patients. This work led to the identification of isolates with large genomic deletions and shows that while these mutations cause a replication defect in the virus, similar mutations have appeared multiple times, independently in the evolution of SARS-CoV-2. Finally, we show that type III CRISPR-Cas systems can be repurposed for molecular detection of SARS-CoV-2 and investigate how these new diagnostic platforms can be improved.

Correspondence: Tanner.wiegand@post.harvard.edu

PREFACE AND OUTLINE

Viruses are ubiquitous genetic parasites that exploit cellular resources in all domains of life. Bacteriophages (i.e., the viruses that infect bacteria), in particular, are thought to be the most abundant biological entities on the planet, with an estimated 10^{31} phage particles existing at any given moment. To defend against these viral predators, about 50% of bacteria and 90% of archaea utilize adaptive immune systems composed of clustered regularly interspaced short palindromic repeats (CRISPRs) and their accompanying Cas proteins.

The function of CRISPR loci were first appreciated by the Spanish scientist Francisco Mojica. In 2005, Mojica noticed that the sequences interspacing CRISPRs were also found in the genomes of bacteriophages. What's more, bacteria were immune to phage infection if one or more of their CRISPR spacers matched a sequence in the genome of the phage. Based on these observations, Mojica and his colleagues hypothesized that CRISPR loci were components of an anti-phage immune system. Similar hypotheses concerning the function of CRISPR loci were independently reported that same year by Pourcel *et al*. and Bolotin *et al*. Just two years later, Rodolphe Barrangou and Philippe Horvath showed that strains of *Streptococcus thermophilus*, which are frequently used for fermentation in the dairy industry, could be immunized against phage infection. Their experiments showed that bacteria that survived the infection acquired a new CRISPR spacer sequence matching part of the phage's genome. These and numerous other researchers would go on to elucidate many of the mechanisms of CRISPR-mediated immunity, illuminating the diversity of Cas components found in these systems.

While the current classification defines 37 subtypes of CRISPR-Cas systems, encompassing a large diversity of components and mechanisms of immunity, all CRISPR-

mediated defense systems operate in three universally conserved stages. These immune systems adapt to new threats by integrating short DNA sequences derived from the invader into CRISPR loci. During CRISPR-RNA (crRNA) biogenesis, CRISPR loci are transcribed and form crRNA-guided surveillance complexes by assembling with Cas proteins. These ribonucleoprotein complexes then patrol the intracellular environment in search of DNA or RNA sequences that are complementary to the spacer sequence of their crRNA-guide. When a target sequence is recognized, cis- or trans-acting nucleases degrade the bound nucleic acid, interfering with phage replication. Given the effectiveness of CRISPR-mediated defense systems, many bacteriophages have evolved anti-CRISPR proteins, which prevent destruction of their genetic material via interactions with CRISPR-Cas components.

At the outset of my doctoral work, it was known that Cas1 and Cas2 proteins were required for the integration of new spacer sequences, but many of the mechanistic details of adaptation remained mysterious. In particular, it was unknown how new spacer sequences were selected and acquired in the model I-F CRISPR system of *Pseudomonas aeruginosa*. High throughput sequencing of spacers acquired in a heterologous *E. coli* expression system have enabled us to understand which sequences are recognized as antigenic and my work suggests that the I-F crRNA-guided surveillance complex participates in the acquisition of new spacer sequences. *In vitro* work and bioinformatic comparisons of CRISPR loci also revealed that new sequence adaptation is strongly dependent on the presence and phasing (i.e., spacing) of sequence motifs in CRISPR leader elements. This work is detailed in chapters one and two of this dissertation.

Later work in the course of my PhD focused on other topics in the CRISPR field, including phage-encoded anti-CRISPR (Acr) proteins. Discovery of the first Acrs were reported by Joe Bondy-Denomy and April Pawluk, who were students of Alan Davidson and Karen Maxwell, in 2013. The repertoire of known Acrs quickly expanded as other groups joined in the search, and today we know of Acrs that inhibit CRISPR-mediated interference in type I, II, III, V, and VI systems. These small proteins use numerous strategies to overcome immune clearance, including stoichiometric inhibitors that bind and sterically-block conformational changes in Cas proteins, and enzymatic inhibitors that covalently modify CRISPR-Cas components. Although Acrs are incredibly diverse and most have novel folds, structures of these proteins reveal that many Acrs use convergent strategies to suppress CRISPR systems. The strategies of Acr-mediated inhibition, and their benefits and drawbacks, are discussed in chapter three. Chapters four and five of this dissertation go on to highlight other recent in advances in the field of CRISPR research, including the discovery of Tn7-like transposons that have coopted CRISPR systems for their own crRNA-guided transposition purposes, and experiments revealing target search mechanisms of the type II surveillance complex (i.e., Cas9).

In late December 2019, government officials in the People's Republic of China announced an epidemic of pneumonia that would go to on reshape not only world history, but also the course of my doctoral work. The etiologic agent of these infections was soon discovered to be a plus-sense single-stranded RNA coronavirus—now called SARS-CoV-2—that quickly spread around the world causing a pandemic of COVID-19 (Coronavirus Disease 2019). As businesses and academic institutions closed down, due to government-mandated social distancing policies meant to curb the spread of SARS-CoV-2, it became difficult to conduct

research that was unrelated to COVID-19. Much like the CRISPR loci I had spent two years studying, my research directives adapted to encompass this new virus. Chapters six through eight of this dissertation detail our contributions to the field of SARS-CoV-2 research and explore the new and exciting field of CRISPR-based viral diagnostics. We further explore this latter topic in chapters nine and ten, by investigating the functional and phylogenetic diversity of two Cas proteins (i.e., Cas10 and Cas7) involved in type III CRISPR-based SARS-CoV-2 detection platforms.

Chapter 1 presents results from *in vivo* adaptation experiments in the I-F CRISPR system of *Pseudomonas aeruginosa.* We show that Cas1 and Cas2 (which is genetically fused to Cas3, a transacting nuclease-helicase, in this system) are necessary and sufficient for inefficient spacer acquisition. The I-F crRNA-guided surveillance (i.e., Csy) complex—which was previously reported to be involved in new sequence adaptation—greatly enhances the efficiency of spacer acquisition. While the exact mechanism of Csy-mediated acquisition remains elusive, we show that mutating a Csy residue required for interactions with sequence motifs that flank *bona fide* targets (i.e., protospacer adjacent motifs) greatly reduces the quantity of newly acquired spacers. We also used a sequencing approach to better understand the mechanisms of I-F spacer selection. High throughput sequencing of CRISPRs with newly acquired spacers indicates that some sequences from the termini and origins of plasmids may be recognized as antigenic, since they are preferentially and reproducibly acquired as new spacers.

Chapter 2 investigates the distribution and phasing of sequence motifs in CRISPR leaders and shows that these motifs are critical to new sequence adaptation. Using bioinformatic and phylogenetic approaches, we show that over half of CRISPR loci have leaders with sequence

motifs that are recognized by integration host factor (IHF) proteins. These leaders additionally contain upstream sequence motifs that had not been previously reported. In the I-E system of *Escherichia coli*, IHF bends the leader DNA into a horseshoe shape during adaptation, facilitating the recruitment of upstream sequence motifs that are recognized by Cas1. Since the spacing of leader motifs are conserved, we hypothesized that the phasing of upstream leader motifs would influence how these motifs are presented to and recognized by CRISPR-associated adaptation proteins. *In vitro* spacer integration assays support these predictions, since perturbations to the conserved spacing between CRISPR leader motifs greatly reduces the ability of Cas1 to integrate new spacers.

Chapter 3 discusses the strategies bacteriophages and other mobile genetic elements employ to subvert CRISPR-mediated interference. In particular, this text focuses on anti-CRISPR (Acr) proteins, many of which bind to Cas proteins to sterically inhibit their immunological activities. We use structures of Acr proteins to highlight common strategies of Acr-mediated immune inhibition, including Acrs that mimic the shape and charge of target nucleic acids, and Acrs that dimerize Cas-proteins to prevent their catalytic activities. We also focus on an emerging class of Acr proteins that function enzymatically, either by covalently modifying the crRNA-guided surveillance complex, or by cleaving second messenger molecules that are involved in immune-signaling.

Chapter 4 discusses the recently discovered phenomenon of CRISPR-associated transposons (CASTs). CASTs are selfish genetic elements that repurpose the components of type I or type V CRISPR systems for RNA-guided transposition of their own genetic cargo. In

particular, we highlight recently determined structures of a I-F CAST complex and discuss what these structures teach us about crRNA-guided transposition of CAST systems.

Chapter 5 discusses the mechanisms of target search by the type II crRNA-guided surveillance complex (i.e., Cas9). Cas9 has been widely adopted for genome-engineering, since it produces crRNA-guided dsDNA breaks in target sequences. Like many surveillance complexes, Cas9 distinguishes between *bona fide* targets and spacer sequences in CRISPR loci by detecting protospacer adjacent motifs (i.e., PAM), short sequences that are absent from the repeats flanking CRISPR spacers. In addition to preventing autoimmune targeting, PAM-dependent search mechanisms accelerate target surveillance. We summarize recent advances in our understanding of how PAMs effect the Cas9 target-search process and discuss how PAM densities are likely to contribute to differential outcomes in editing of bacterial and human genomes.

Chapter 6 pivots from CRISPR-related research and presents data obtained from monitoring SARS-CoV-2 RNA levels in municipal wastewater. Using reverse transcription quantitative PCR (RT-qPCR) on sewage samples, we show that spikes in viral RNA precede increases in hospital-reported positive tests for COVID-19 by two to four days. Additionally. we used long-read sequencing to assemble a SARS-CoV-2 genome from wastewater and show that the dominant viral genotype circulating in Bozeman, Montana in late March 2020 was most closely related to viral isolates from California, Australia, and New York. Collectively, this work shows that surveillance of municipal wastewater is the best early indicator of SARS-CoV-2 prevalence and genotypic diversity in a community.

Chapter 7 continues on the topic of SARS-CoV-2, presenting results that reveal how mutations to ORF7a effect viral replication dynamics. Genetic surveillance of SARS-CoV-2 circulating in the Montana community of Bozeman, led to the identification of seven isolates that had a 115-nucleotide deletion in the gene encoding for accessory protein ORF7a (i.e. ORF7a$^{\Delta 115}$). This mutation causes a frameshift and the insertion of a premature stop codon, truncating the latter two thirds of ORF7a. Analyses of publicly available SARS-CoV-2 genomes show that while this particular mutation hadn't been previously reported, many other SARS-CoV-2 genomes have mutations that would produce large deletions or truncations in ORF7a. Since ORF7a has been implicated in the modulation of the human system, we hypothesized that truncations of ORF7a would result in a viral replication defect. Infection assays in human and monkey cell lines corroborate this prediction, showing that SARS-CoV-2 with ORF7a$^{\Delta 115}$ produce significantly lower titers than wildtype virus. RT-qPCR performed on cells infected with mutant or wildtype virus reveals elevated levels of type I interferon when ORF7a is truncated, reinforcing previous work indicating ORF7a is involved in immune inhibition.

Chapter 8 marries the fields of SARS-CoV-2 and CRISPR-Cas research, presenting results that show type III CRISPR defense components can be repurposed for the molecular detection of viral RNA. Limited access to testing resources created widespread concern during the first wave of the COVID-19 pandemic. To address this issue, numerous groups developed type V and VI CRISPR-based diagnostic platforms for rapid and specific detection of SARS-CoV-2 infection. We hypothesized that type III CRISPR systems could also be repurposed as diagnostics, since their natural targets are viral RNA sequences and they amplify signals of infection by producing second messenger molecules. Results from an *in vitro* workflow show

that the type III CRISPR system of *Thermus thermophilus* can be commandeered for sequence-specific detection of SARS-CoV-2. This assay produces visible readouts when samples contain viral RNA at clinically relevant levels (i.e., 10^7-10^8 viral RNA copies).

Chapter 9 highlights efforts to improve type III CRISPR-based diagnostics by broadening our understanding of the functional and phylogenetic diversity of Cas10 polymerases. Cas10 is the large subunit of type III crRNA-guided surveillance (i.e., Csm/Cmr) complexes, and produces cyclic oligoadenylate (cOA) molecules from ATP when the Csm/Cmr complex binds a target. Using a bioinformatic pipeline, we identify 8,088 Cas10 proteins in genomic and metagenomic databases. We show that active site motifs required for polymerase activity are not universally conserved in all Cas10 proteins. *In vitro* activity assays with five phylogenetically distant Cas10 proteins additionally reveal that all of the crRNA-guided surveillance complex components and a complementary RNA target are required for Cas10-mediated cOA synthesis. Surprisingly, this work reveals that all five Cas10 proteins we tested have DNA-dependent DNA (DdDp) polymerase activity. While the mechanistic details of this DdDp activity remain elusive, our work reveals Cas10 proteins that have been genetically fused to components of innate immune systems or toxin-antitoxin (TA) systems, implying that Cas10 proteins can acquire new functions that may be of biological or biotechnological interest.

Chapter 10 investigates the phylogenetic diversity of Cas7 proteins and shows that—like Cas10 proteins—some Cas7 proteins have been genetically fused to modules of innate immune and TA systems. Cas7 proteins are "backbone" subunits of type I, III, and IV crRNA guided surveillance complexes, and are accordingly diverse. This work takes a similar approach to work detailed in Chapter 9, and we identify 22,983 Cas7 proteins in genomic and metagenomic

databases. In particular, we highlight novel genetic fusions of Cas7 proteins and predict how these fusions impact crRNA-guided interference.

CHAPTER ONE

REPRODUCIBLE ANTIGEN RECOGNITION BY THE TYPE I-F CRISPR SYSTEM

Contribution of Authors and Co-Authors

Manuscript in Chapter 1

Author: Tanner Wiegand

Contributions: designed the research, performed the research, analyzed data, wrote the manuscript

Co-Author: Ekaterina Semenova

Contributions: designed the research, performed the research

Co-Author: Anna Shiriaeva

Contributions: designed tools for data analysis

Co-Author: Ivan Fedorov

Contributions: performed the research

Co-Author: Kirill Datsenko

Contributions: engineered bacterial strain for experiments

Co-Author: Konstantin Severinov

Contributions: designed the research

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

Tanner Wiegand, Ekaterina Semenova, Anna Shiriaeva, Ivan Fedorov, Kirill Datsenko, Konstantin Severinov, and Blake Wiedenheft

The CRISPR Journal

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Mary Ann Liebert, Inc.
Volume 3
Issue 5
20 October 2020
DOI: 10.1089/crispr.2020.0069

REPRODUCIBLE ANTIGEN RECOGNITION BY THE TYPE I-F CRISPR SYSTEM

Tanner Wiegand,[1] Ekaterina Semenova,[2] Anna Shiriaeva,[3] Ivan Fedorov,[3] Kirill Datsenko,[2] Konstantin Severinov,[2,3,4,5] Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Waksman Institute of Microbiology, Rutgers, the State University of New Jersey, Piscataway, NJ. [3]Skolkovo Institute of Science and Technology, Skolkovo, Russia. [4]Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia. [5]Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia.

Abstract

CRISPR-associated proteins 1 and 2 (Cas1–2) are necessary and sufficient for new spacer acquisition in some CRISPR-Cas systems (e.g., type I-E), but adaptation in other systems (e.g., type II-A) involves the crRNA-guided surveillance complex. Here we show that the type I-F Cas1–2/3 proteins are necessary and sufficient to produce low levels of spacer acquisition, but the presence of the type I-F crRNA-guided surveillance complex (Csy) improves the efficiency of adaptation and significantly increases the fidelity of protospacer adjacent motif selection. Sequences selected for integration are preferentially derived from specific regions of extrachromosomal DNA, and patterns of spacer selection are highly reproducible between independent biological replicates. This work helps define the role of the Csy complex in I-F adaptation and reveals that actively replicating mobile genetic elements have antigenic signatures that facilitate their integration during CRISPR adaptation.

Introduction

CRISPR loci and their associated *cas* genes are conserved components of adaptive immune systems that protect bacteria and archaea from foreign genetic elements.[1,2] These immune systems are phylogenetically and functionally diverse, but all acquire immunity by preferentially integrating short fragments of foreign DNA at one end of a CRISPR locus.[3-6] Conserved CRISPR- associated proteins 1 and 2 (i.e., Cas1 and Cas2) have been shown to be necessary and sufficient for integrating foreign DNA into the CRISPR loci in some system subtypes (e.g., I-E),[7] while in vivo adaption in other CRISPR subtypes requires additional proteins (e.g., Cas4, Csn2, Cas9).[8-11]

The type II-A crRNA-guided surveillance complex (i.e., Cas9) facilitates integration of sequences flanked by a protospacer adjacent motif (PAM),[11,12] and chimeric Cas9 proteins have been used to demonstrate that the PAM interacting domain is directly involved in this process.[10] Similarly, the type I-F crRNA-guided surveillance (Csy) complex has been implicated in naïve sequence adaptation,[13] but the role played by this complex is yet to be determined.

Based on the previous work done in type II systems showing that Cas9 plays a necessary role in selecting sequences for integration (i.e., prespacers), we hypothesized that the type I-F crRNA-guided surveillance complex (i.e., Csy or I-F Cascade complex) has a similar role in I-F CRISPR adaptation. Here, we show that the Cas1 and Cas2/3 fusion proteins (Cas1–2/3), which form an integration complex in I-F systems,[14-16] are capable of selecting PAM containing prespacers for inefficient adaptation independent of the Csy complex. However, the presence of the Csy complex increases the fidelity of PAM selection and enhances the efficiency of naive CRISPR adaptation. Moreover, our work reveals that protospacers are not selected at random;

instead specific protospacers are consistently selected from specific regions of extrachromosomal DNA. Persistent protospacer preferences indicate that some DNA sequences may be recognized as antigenic and this work helps explain how CRISPR systems avoid autoimmunity.

Materials and Methods

Adaptation assays in *Escherichia coli*

Spacer acquisition assays were performed using an engineered strain of *Escherichia coli* BL21-AI that contains a type I-F CRISPR array (KD740) in place of the I-E CRISPR locus. These cells are identical to those de- scribed by Vorontsova et al.,[13] but an orphan type I-F CRISPR locus has been removed using Red recombinase.[17,18] The I-F CRISPR contains a single-spacer sequence (5'-ACGCAGTTGCTGAGTGTGATCGATGCCATCAG-3') that targets the J protein of phage lambda and should not trigger priming, since this spacer has no homology to the chromosome or expression plasmids used in these studies.[13,19,20] The I-F CRISPR of KD740 is flanked by a 134 bp leader sequence. The CRISPR is transcribed from LacI-repressed T7 RNA polymerase promoter.

The *cas1–2/3* genes from *Pseudomonas aeruginosa* PA14 were cloned into a spectinomycin resistance (2S) LIC vector with a ColE1 origin of replication. The *cas1–2/3* genes are cloned downstream of a T7 promoter. The construction of this plasmid (pCas1–2/3) has been previously described,[15] and the plasmid has been deposited on Addgene (#89240). pCsy was constructed by cloning genes coding for the Csy complex from PA14 (i.e., *cas8f, cas5f, cas7f, cas6f*) downstream of a T7 promoter in pRSF-1 and was deposited on Addgene (#89232). Q5 mutagenesis was used to introduce a mutation in *cas8f* (N250A) that prevents Cas8f- mediated

PAM-recognition (primers P-Csy$_{PAMmut}$-F and P-Csy$_{PAMmut}$-R in Supplementary Table S1.1),[21] resulting in the pCsy$_{PAM}$ expression vector.

KD740 cells were transformed by electroporation with pCas1–2/3, and either the WT Csy expression vector (pCsy) or the PAM-recognition-deficient Csy mutant (pCsy$_{PAM}$). Transformants were selected on Luria-Bertani (LB) agar plates containing 50 μg/mL spectinomycin (pCas1–2/3) and 50 μg/mL kanamycin (pCsy/pCsy$_{PAM}$), but no antibiotics were used in downstream adaptation experiments. Cells were grown at 32°C in LB media supplemented with 1 mM L-arabinose and 1 mM IPTG for a total of 72 h in a shaking incubator at 225 RPMs. Cultures were diluted (1:100) in fresh media every 24 h.

Spacer acquisition was detected using "internal" and "degenerate" primers with Q5 polymerase (NEB) via the recently described CAPTURE method.[22] Briefly, 1μL of liquid culture was used as template for polymerase chain reaction (PCR) with "internal" primers (i.e., P1 and P2) (Supplementary Table S1.1), and products were visualized on 1.5% agarose gels. PCR products were extracted from gels and purified using Zymoclean Gel DNA Recovery Kits (Zymogen) or the BluePippin size selection system (Sage Science) with a "Range" setting of 220–300 bp on a 2% agarose cassette. These size selected oligos were then used as templates for a subsequent round of PCR with "degenerate" primers (i.e., P3 and P2) (Supplementary Table S1.1), and products were subsequently visualized on 1.5% agarose gels. These experiments did not require approval by an IRB.

<u>High-throughput sequencing</u>

Products from the second PCR amplification were extracted from agarose gels and were purified using a Zymoclean Gel DNA Recovery Kit (Zymogen). These products were used as

templates for PCR with Q5 polymerase (NEB) using barcoded primers P-HTS-$F_{1–5}$ and P-HTS-R (Supplementary Table S1.1). Samples were sequenced with 1 x 150 bp single-end runs on an Illumina NextSeq and included two independent biological replicates of KD740 with pCsy and pCas1–2/3, and three independent biological replicates of KD740 with pCas1–2/3.

Spacer sequence analysis and statistics

High-throughput sequencing (HTS) data were processed with the ShortRead package from Bioconductor.[23] Sequencing reads contained one, two, or three CRISPR, representing unexpanded arrays (i.e., no adaptation), arrays with one newly acquired spacer, and arrays with two newly acquired spacers, respectively. Only reads corresponding to one acquisition event were examined for downstream analyses, to prevent the analysis of spacers acquired through priming. Sequences between CRISPR (i.e., spacers) were extracted. Spacer sequences corresponding to the original WT spacer present in KD740 were removed.

Extracted spacers were aligned to the *E. coli* genome (NC_012947.1) and *cas* gene expression plasmids present in each sample. Sequences were aligned using the Biostrings package, allowing for one mismatch.[24] Spacer sequences that align with pCas1–2/3 were used for the analysis shown in Figure 3B and C. To assess spacers that uniquely align to pCas1–2/3 (Figs. 2 and 3D, E), we discarded sequences that produced more than one alignment or that aligned to a 565 bp region shared by the *ori* sequences of pCas1–2/3 and pCsy (94% nucleotide identity).

Pearson correlation coefficients were calculated by comparing the quantities of uniquely aligned protospacers that were present at each position of pCas1–2/3 using the ggpubr package,[25] and graphs of protospacer positions and quantities were rendered with the ggplot2 package.[26] A

summary of the HTS results can be found in Supplementary Table S3 and spacer sequences extracted from sequencing reads are included in Supplementary Table S4.

Results

Csy-PAM interactions increase I-F CRISPR adaptation efficiency

To determine if Csy-PAM interactions are required for I-F spacer acquisition, we screened for adaptation using E. coli KD740 cells overexpressing Cas1–2/3 (I-$F_{Cas1–2/3}$) and either the wild-type Csy complex (I-$F_{Cas+Csy}$) or a Csy mutant (I-$F_{Cas+CsyPAMmut}$) with reduced PAM recognition (Fig. 1A).[21] Adaptation assays were performed by PCR amplification of the leader end of the I-F CRISPR array (Fig. 1B). PCR performed on cells expressing both Cas1–2/3 and the Csy complex (I-$F_{Cas+Csy}$) results in two clearly visible PCR products. The smaller product corresponds to the expected size of the wild-type, unexpanded CRISPR (CR_{WT}), and the larger product corresponds to the addition of one spacer-repeat unit. In contrast, PCR performed on cells lacking the Csy complex (I-$F_{Cas1–2/3}$) or expressing the PAM sensing mutant (I-$F_{Cas+CsyPAMmut}$) resulted in only the CR_{WT} product, in agreement with earlier data.[13] These results suggest that Csy and more specifically Csy-PAM interactions are required for efficient I-F adaptation.

Cas1–2/3 are necessary and sufficient for inefficient I-F spacer acquisition

The observation that wild-type Csy is required for efficient spacer acquisition *in vivo* is consistent with previous results,[13] but *in vitro* experiments have shown that Cas1–2/3 are sufficient for spacer processing and integration.[16] We wondered if *in vivo* adaptation occurs

without the Csy complex, but at levels below the limit of detection of the standard PCR-based assay.

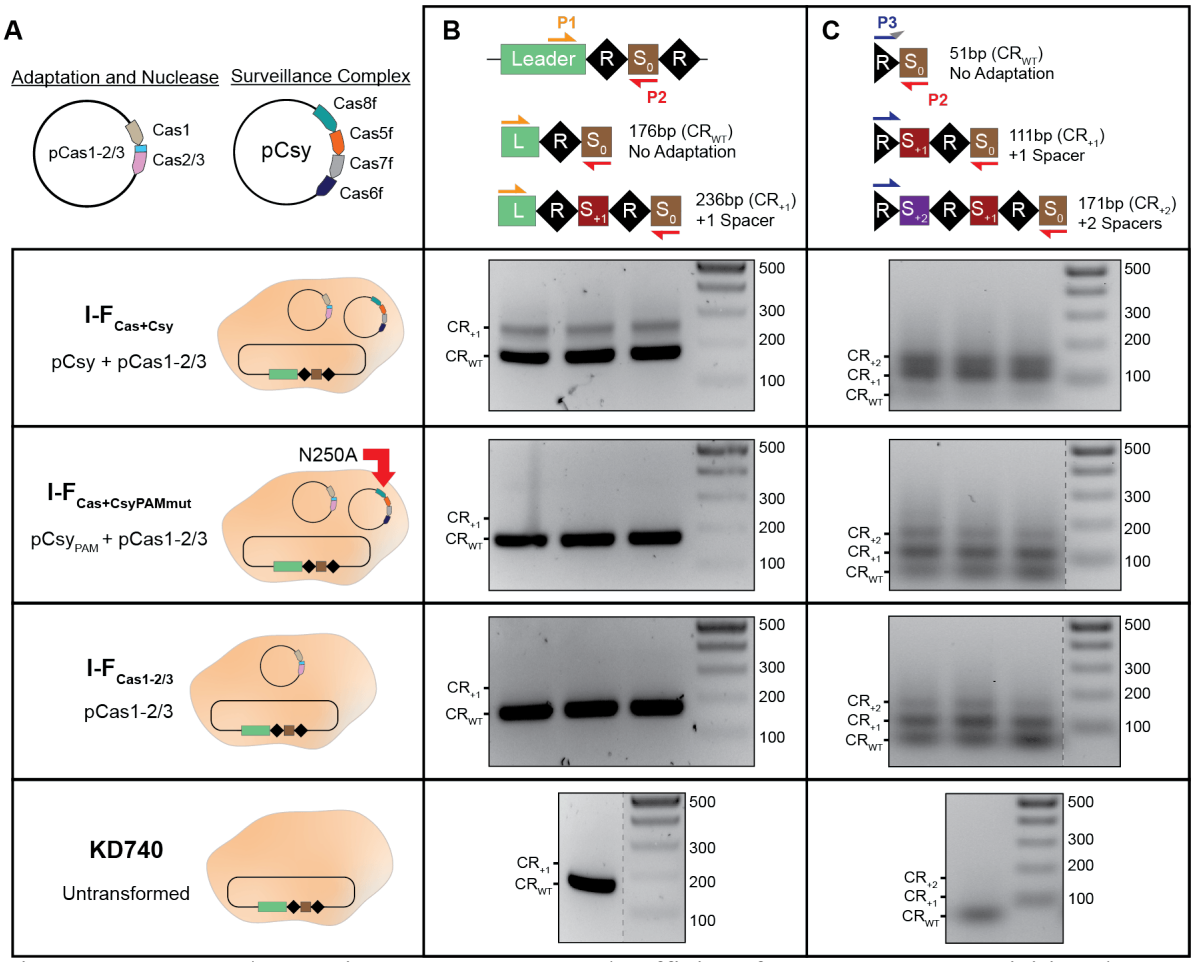


Figure 1.1. Cas1–2/3 proteins are necessary and sufficient for new spacer acquisition, but Csy accelerates adaptation. (A) Schematics of the Cas1–2/3 and Csy subunit (Cas8f, Cas5f, Cas7f, and Cas6f) expression vectors are shown above schematics for each of the conditions tested. *Escherichia coli* KD740 cells are shown as tan ovals. (B) PCR amplification with ‘‘internal’’ (P1 and P2) primers to detect adaptation events. (S0 is initial spacer that targets phage λ.) Expected sizes of PCR products for CRISPR with acquired spacers (CR_{+1} and CR_{+2}) or no acquisition (CR_{WT}) are labeled. Adaptation events detected after PCR with primers P1 and P2. Each lane represents a biological replicate. (C) PCR products from (B) were enriched for higher molecular weight species, and then used as templates for a subsequent round of CAPTURE PCR with primers P3 and P2. Products corresponding to expanded arrays are visible in all biological replicates for I-$F_{Cas+Csy}$, I-$F_{Cas+CsyPAMmut}$, and I-$F_{Cas1–2/3}$. Cas, CRISPR-associated protein; PAM, protospacer adjacent motif; PCR, polymerase chain reaction.

To detect rare adaptation events in I-$F_{Cas+CsyPAMmut}$ and I-$F_{Cas1–2/3}$ samples, we cut out regions of the gel that would contain higher molecular weight PCR products corresponding to expanded arrays and repeated the PCR according to a recently published CAPTURE PCR protocol.22 Size-enriched PCR products were amplified using a forward primer with a degenerate 3' nucleotide that anneals in the CRISPR (P3) and a reverse primer that is complementary to the phage lambda (λ)-targeting spacer (P2) (Fig. 1C). This approach enables detection of rare (i.e., 1 in 10^5 cells) adaptation events, and is two orders of magnitude more sensitive than the standard PCR protocol.[22]

Using the CAPTURE primer set (P2 and P3), PCR products corresponding to expanded arrays (111 and 171 bp) were detected in samples from all three conditions (i.e., I-$F_{Cas+Csy}$, I-$F_{Cas+CsyPAMmut}$, and I-$F_{Cas1–2/3}$ cultures) (Fig. 1C). Notably, strong bands corresponding to one (111 bp) and two (171 bp) new spacer-repeat units were visible in cells expressing both Cas1–2/3 and Csy (I-$F_{Cas+Csy}$), suggesting that size selection of the PCR template successfully enriches expanded arrays (Fig. 1C). These results indicate that Cas1–2/3 proteins are necessary and sufficient for I-F spacer acquisition, but adaptation in this system is inefficient in the absence of the wild-type Csy complex.

The Csy complex increases the percentage of PAM-proximal spacers

Cas9 plays a critical role in prespacer selection during *in vivo* adaptation of type II CRISPR systems.[10,11] While the mechanistic details of this process remain obscure, we do know that Cas9 has to be loaded with a tracrRNA and crRNA-guide and that the PAM sensing residues play a critical role in determining the sequence of newly acquired spacers.[10,11] To determine if Csy plays a similar role in type I-F adaptation, we used Illumina NextSeq to sequence higher

molecular weight CAPTURE PCR products from cells expressing Cas1–2/3 alone (I-$F_{Cas1–2/3}$) or Cas1–2/3 and Csy (I-$_{FCas+Csy}$), then protospacer flanking sequences were analyzed for the presence of PAMs. We focused this analysis on 32bp spacers that uniquely aligned to a single protospacer target on the pCas1–2/3 plasmid, since this represented the largest pool of uniquely aligned spacers in both I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ samples.

In total, we analyzed 227,834 and 205,354 sequences flanking I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ protospacers, respectively (Fig. 2A). In cells expressing Csy and Cas1–2/3, the majority (98.3–98.9%) of protospacers were flanked by a PAM (Fig. 2B). This trend was consistent across biological replicates, and these numbers agree with previous work showing high-fidelity PAM selection (97.8%) in this system.[13] While the majority of protospacers acquired in cells expressing Cas1–2/3 alone (I-$F_{Cas1–2/3}$) were most often flanked by a PAM (87.9–88.6%), the occurrence of a PAM in these samples (i.e., Cas1–2/3 alone) was significantly less than what was measured for cells expressing Csy and Cas1–2/3 (T-test; $p = 1.2 \times 10^{-3}$) (Fig. 2B). These differences indicate that Cas1–2/3 are capable of acquiring spacers flanked by a PAM, but Csy significantly increases the fidelity of PAM selection.

Previous work has shown that newly acquired spacers are often flanked by incorrect PAMs due to slipping of Cas1–2/3 during prespacer capture or processing.[20,27,28] To determine if uniquely aligned protospacers flanked by an incorrect PAM could have been produced by slipping, we searched for a PAM three nucleotides upstream and downstream of each protospacer (Fig. 2A).

For the three I-$F_{Cas1–2/3}$ biological replicates we sequenced, 10.6–11.4% of the 32 bp protospacers that were not directly flanked by a PAM were within three nucleotides of a PAM,

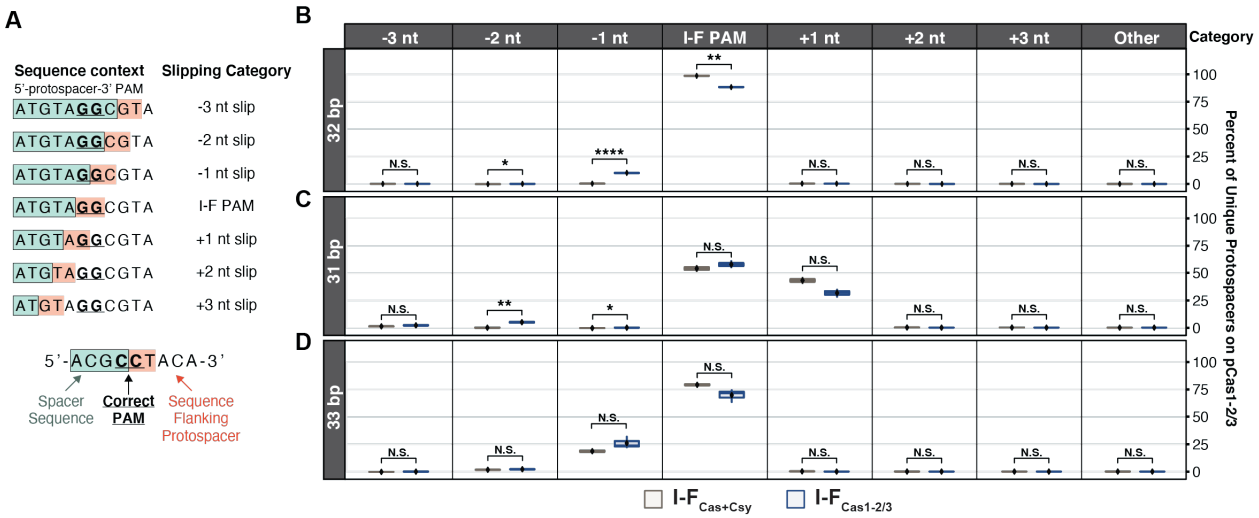


Fig 1.2. Cas1-2/3 acquire mostly PAM-flanked spacers, but Csy increases the fidelity of PAM selection. (A) Sequence context of protospacers for spacers acquired through Cas1–2/3 slipping. (B–D) Percentages of 31, 32, and 33 bp protospacers that uniquely map to pCas1–2/3 and were acquired through different categories of Cas1-2/3 slipping. Spacers that produced more than one alignment to the genome or *cas* gene expression plasmids were not included in this analysis. Significances of T-tests performed for biological replicates between each condition are denoted as asterisks (N.S.: $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ****$p \leq 0.0001$).

suggesting that Cas1–2/3 slipping events and subsequent incorrect processing at the PAM-spacer boundary may be responsible for the integration of most spacers without a PAM (Fig. 2B). In particular, I-$F_{Cas1–2/3}$ spacers that were acquired with a slip of -1nt accounted for most (92.2–93.5%) of the 32bp protospacers that were not flanked by a PAM and represented a large portion (9.9–10.7%) of the uniquely aligned I-$F_{Cas1–2/3}$ spacer population. This represents a significant increase in -1 nt slipped spacers for I-$F_{Cas1–2/3}$ samples over I-$F_{Cas+Csy}$ (0.2–0.6%) samples ($p = 9.9 \times 10^{-5}$; T-test) (Fig. 2C).

Collectively, these data indicate that Cas1–2/3 mediate prespacer trimming at the PAM-spacer boundary with about 90% accuracy in vivo, but the fidelity of this processing is increased by the presence of the Csy complex.

Since previous work has shown that aberrantly sized spacers are often associated with Cas1–2/3 slipping,[20] we examined the PAMs of 31 and 33bp spacers that uniquely align to pCas1–2/3. I-$F_{Cas1–2/3}$ samples had significantly more 31 bp spacers that were acquired with a slip of -1 or -2nt compared with I-$F_{Cas+Csy}$ samples ($p = 1.6 \times 10^{-2}$ and $p = 3.1 \times 10^{-3}$, respectively; T-test) (Fig. 2C). As previously reported,[20] slipping of +1nt was elevated in 31bp spacers acquired in both conditions compared with +1 nt slipping for 32 bp spacers (Fig. 2B, C). Thirty-three base pair spacers had more -1nt slipping overall, and although this trend was more pronounced in I-$F_{Cas1–2/3}$ spacers than in I-$F_{Cas+Csy}$ spacers, the difference between these two conditions was not significant ($p = 0.123$; T-test) (Fig. 2D). Considered together, these data are consistent with the observation that slipping occurs more frequently when the Csy complex is not present.

Spacers are preferentially acquired from the terminus of plasmid replication

The origin of plasmid replication initiation (*ori*) has previously been shown to be a "hotspot" for spacer acquisition,[6,13,20] however, it is unclear if the surveillance complex is involved in the spacer selection process. To determine if the Csy complex effects the frequency or location of spacer selection, we calculated the number of newly acquired spacers derived from the *E. coli* genome and the *cas* gene expression plasmids. The majority of spacers (44.7–45.8%) acquired in I-$F_{Cas+Csy}$ cells were homologous to sequences that are identical in both expression plasmids (i.e., pCas1–2/3 and pCsy), and most of these spacers (97.3–98.3%) mapped to discrete locations near the origin of replication initiation (*ori*) (Fig. 3A, B). This preference for spacers derived from the *ori* occurred independent of Csy expression, since 69.6–77.1% of spacers acquired in I-$F_{Cas1–2/3}$ samples also mapped to the *ori* (Fig. 3A, C). These patterns suggest that the *ori* is a hotspot for acquisition, regardless of whether or not Csy is present.

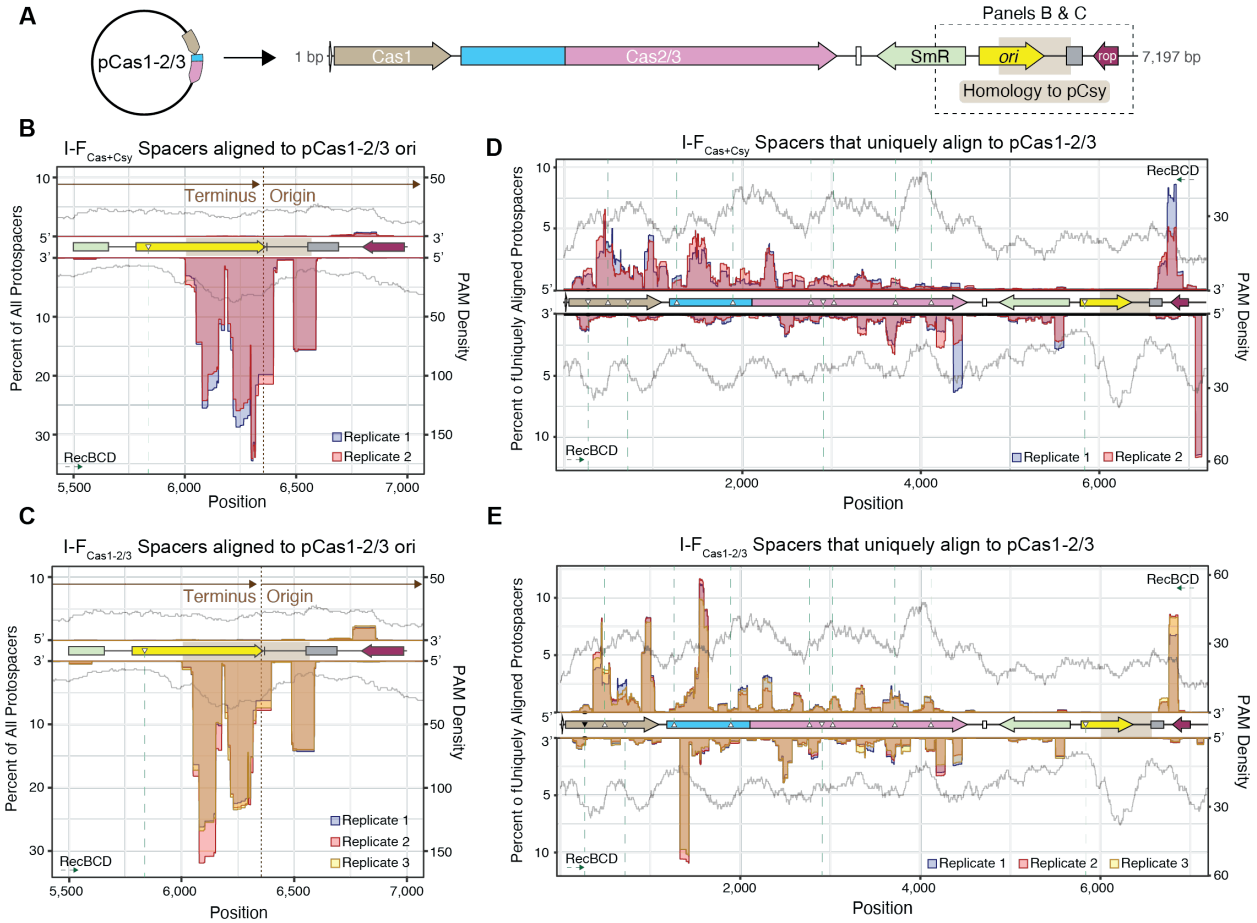


Fig. 1.3. Protospacer hotspots are reproducible across multiple biological replicates. (A) Linearized genetic map of Cas1–2/3 expression plasmid. Regions shown in (B, C) are marked with dotted outline. (B, C) Positions and quantities of all protospacers that align to *ori* of pCas1–2/3, including spacers that produce alignments to additional sequences (e.g., pCsy). Protospacer locations are mapped using a 50 nt sliding window. Data from each biological replicate are represented as a different color. PAM densities (gray) mapped with a 150 nt sliding window. Direction and start/end positions of plasmid replication shown in brown. Consensus Chi sites are shown as triangles filled in with black, and Chi sites with a single base mismatch are shown as triangles with a white fill. Dotted green lines indicate the boundaries of Chi sites, and RecBCD directions of cleavage are shown for each strand of DNA. (D, E) Positions and quantities of protospacers that produce a single unique alignment to pCas1–2/3 are shown as in (B, C).

The ColE1 *ori* of pCas1–2/3 shares ~565 bp of homology with the RSF1030 *ori* of pCsy (94% nucleotide identity), and spans both the start and end of replication for these unidirectionally replicating plasmids.[29] Most spacers acquired from this region are derived from sequences near the terminus of replication, rather than from replication initiation (Fig. 3B, C).

Replication termini of bidirectionally replicating genomes and plasmids have previously been shown to be hotspots of spacer acquisition,[6,13] and these data suggest that termini of unidirectionally replicating plasmids present similar antigenic signals that increase the number of spacers acquired from these regions.

Preference for terminus-centered acquisition have previously been explained by the action of RecBCD, which produces prespacer substrates during repair of dsDNA breaks (DSBs).[6] DSBs occur more often near the terminus of replication where replication forks stall,[6] and RecBCD nucleases degrade both strands of linear dsDNA until RecBCD reaches an asymmetric, octameric Chi site (5'-GCTGGTGG-3').[30] pCas1–2/3 and pCsy both contain a single imperfect Chi site (5'-GCTGGTAG-3'; mutation underlined) that is present near the *ori*. However, previous work has shown that this mutated Chi sequence is seldom recognized by RecBCD (<2% of consensus Chi-site activity),[31,32] and this mutated Chi site is improperly oriented to stop RecBCD degradation, with respect to the terminus of replication (Fig. 3B, C).[30] These data suggest that Chi-site distributions alone do not explain the boundaries of protospacer peaks near the pCas1–2/3 terminus.

Acquisition of spacers from the *E. coli* genome was rare (0.004–0.032%) and percentages of genomic protospacers were not significantly different between biological replicates of I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ (p = 0.15; T-test) (Supplementary Table S3). These results were expected for cells expressing the Csy complex, since the acquisition of "self" spacers would target the Cas2/3 nuclease to the *E. coli* chromosome, resulting in autoimmunity. However, the strong preference for plasmid-derived spacers in cells expressing Cas1–2/3 alone (I-$F_{Cas1–2/3}$) suggests that either RecBCD preferentially generates prespacers from replicating plasmids or that the

Cas1–2/3 integration complex is capable of distinguishing self from nonself during I-F adaptation.

Some spacers are reproducibly acquired

To determine if there are hotspots for protospacer selection beyond the *ori*, we mapped the positions and quantities of spacers outside this region. Specifically, we examined 32bp spacers extracted from I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ reads that uniquely align to pCas1–2/3. Protospacers outside the *ori* mapped primarily to the 5' end of transcriptionally active genes, including *cas1* and *cas2/3*, for both I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ samples (Fig. 3D, E). In contrast, very few spacers were acquired from the intergenic region between *cas2/3* and the spectinomycin resistance gene (*smR*) in either I-$F_{Cas+Csy}$ (0.2–0.4%) or I-$F_{Cas1–2/3}$ (0.0–0.1%) samples (Fig. 3D, E). These data are consistent with previous results suggesting that in addition to replication termini, protospacers are preferentially acquired from highly transcribed regions of foreign DNA.[20]

To determine if specific sequences are preferentially targeted for integration across biological replicates, as previously reported,[6,13,20] we calculated Pearson correlation coefficients for positions and quantities of protospacers unique to pCas1–2/3 (Fig. 3D, E). This analysis revealed that I-$F_{Cas+Csy}$ protospacer profiles from two biological replicates were positively correlated ($R = 0.83$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. S1A). In addition, protospacer profiles from three I-$F_{Cas1–2/3}$ replicates were nearly perfectly correlated ($R = 0.98–0.99$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. S1B). To determine how related the I-$F_{Cas+Csy}$ and I-$F_{Cas1–2/3}$ protospacer profiles were, we calculated mean protospacer quantities from each biological replicate of each

condition. This analysis reveals that the two data sets are positively correlated ($R = 0.78$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. S1C).

Taken together, these data indicate that hotspots of naive spacer acquisition are remarkably consistent with and without the Csy complex. Cas1–2/3, therefore, appears to be the primary determinant of spacer selection during naive I-F sequence adaptation. While we were unable to find similarities in the primary structure of highly acquired spacer sequences, the preponderance of protospacer hotspots in actively transcribed areas of the plasmid suggests that a higher-level DNA structure may be recognized as antigenic by Cas1–2/3.[13,20]

Alternatively, DSBs that form as a result of transcription may lead to increased acquisition from these regions due to the increased availability of RecBCD degradation products.[20] One complete Chi site and nine sites with a single mismatch to the consensus Chi sequence are present on pCas1–2/3. Previous work has shown that Chi sites with a single base mutation are sufficient, although less efficient (<2–38% of consensus Chi activity) for impeding RecBCD degradation of dsDNA.[30-32] While some protospacer peaks occur near the boundaries of these RecBCD-halt signals, not all of the protospacer peaks are constrained by Chi sites as previous experiments in a I-E system have shown (Fig. 3B–E).[6] These data suggest that the I-F integration complex may provide an additional level of prespacer selection specificity during CRISPR adaptation.

Discussion

We have shown that the Cas1–2/3 proteins are necessary and sufficient for low levels of PAM-proximal sequence acquisition in the type I-F CRISPR-Cas system of *P. aeruginosa*. However, it is probably unlikely that spacers are often acquired independent of the surveillance

complex, since adaptation is difficult to detect even when Cas1–2/3 are overexpressed. Our data reveal that the Csy complex plays a role in increasing the efficiency of naive adaptation, and that this process is dependent on Csy-PAM interactions (Fig. 1). These findings differentiate Csy-facilitated adaptation from Cas9-dependent acquisition. In type II systems, PAM-sensing by Cas9 is required for the acquisition of spacers flanked by PAMs, but PAM-sensing mutants of Cas9 still support efficient adaptation of spacers that are not flanked by PAMs.[10]

While the mechanistic details of Csy-mediated enhancement of adaptation remain murky, our results suggest that the Csy complex not only increases the frequency of protospacers flanked by a PAM, but also helps define the PAM-spacer boundary. However, these data and their interpretations may be affected by additional selective pressures. Cells expressing both Csy and Cas1–2/3 can eliminate the *cas* gene expression plasmids via CRISPR-mediated interference. Cells replicating without the metabolic burden associated with supporting plasmid replication may have a growth advantage,[33,34] and this advantage could account for some of the spacer acquisition enhancements that we detect in cells expressing both Cas1–2/3 and Csy. Additional experiments are necessary to clarify the influence of newly acquired spacers on cells during adaptation.

The Cas3 lobe of the genetic Cas2–3 fusion protein may be involved in spacer maturation,[20] since the insertion of a stop codon between Cas2 and Cas3 or inactivating mutations of the SF2 helicase or HD nuclease active sites all prevent efficient spacer acquisition *in vivo*.[13] However, these same mutations do not prevent spacer processing and integration *in vitro*.[16] While our work does not address the role of Cas3 activity in naive spacer acquisition, our

results do demonstrate that cells expressing Cas1–2/3 alone are sufficient for new spacer integration.

Previous work has shown that Cas1–2/3 are capable of processing prespacers at the PAM-boundary with ~50% accuracy in vitro,16 but our results indicate that this process happens with greater accuracy (~90%) in vivo (Fig. 2). These differences may arise from the availability of other cellular nucleases (e.g., DNA PolIII, ExoT, ExoI, or ExoIII) involved in trimming prespacer substrates in the I-E and II-A systems.[35,36] However, the participation of cellular nucleases in I-F spacer maturation remains to be demonstrated, and additional studies are needed to determine if and how Csy-mediated spacer processing occurs.

Strikingly, we demonstrate that spacer sequences acquired in one experiment were frequently acquired in other biological replicates, reinforcing previous findings and suggesting that plasmids present antigenic signatures that prompt preferential adaptation from these elements.[6,13] The vast majority of reproducible protospacers were confined to the origin and terminus of plasmid replication. Higher copy numbers of DNA at origins of replicating genomes have previously been used to explain high levels of spacer acquisition at these loci.[6] In fact, when spacer counts are normalized to total cellular DNA content, protospacers derived from the origin become less pronounced.[6] Similar dynamics may be at play in our data, since regions near the origin of plasmids will be replicated first and will thus have a higher copy number in the cell.

However, the high frequency of spacers derived from plasmid termini is more difficult to explain. Acquisition from the termini occurs reproducibly and independent of the Csy complex (Fig. 3). While we cannot rule out an increase in acquisition at this locus due to prespacer substrate generation by RecBCD DNA repair machinery, our data indicate that Chi-site

distributions alone do not explain the reproducible patterns of protospacer acquisition. These data argue for an additional level of antigen-specific detection by the type I-F Cas1–2/3 integration complex, although the mechanistic details of this process remain to be determined.

Spacers acquired from the genome were rare in our data (0.004–0.032% of total spacers), regardless of whether or not cells expressed the Csy complex. A similar dearth (<1%) of genome-derived spacers has been detected in other I-F overexpression systems.[13] In contrast, the I-F system of *Pectobacterium atrosepticum* has been reported to acquire more genomic spacers (~16%) when *cas* genes are expressed at wild-type levels.[20]

Direct comparison between results of different experiments is, however, difficult due to differing experimental conditions (e.g., endogenous vs. heterologous expression, plasmid copy numbers, and different bacterial species). For example, naive adaptation experiments with I-E CRISPR systems result in the opposite dynamic. Cas1–2 overexpression increases I-E acquisition from the genome, while lower *cas* expression levels produce less genomic spacers.[6] In contrast to both I-E and I-F adaptation, naive spacers acquired in the I-B system of *Pyrococcus furiosus* nearly all come from the genome.[37] These subtype-specific differences may reflect differing mechanisms of Cas-mediated antigen recognition or may simply be the result of different experimental conditions.

While the specific antigenic signatures of plasmid DNA remain nebulous, cellular localization may play a role in recognition of these foreign genetic elements. Cas2 proteins from the I-E system of E. coli have previously been shown to localize to the poles of bacterial cells,[38] and similar polar localization has been reported for plasmids with ColE1 origins of replication.[39] Many bacteriophages also target cellular poles for entry and k phages replicate at cellular poles.[40]

In contrast, the bacterial chromosome typically occupies the cellular midpoint.[41] Collectively, these data suggest that CRISPR- associated adaptation proteins may rely on DNA location as one of the signatures that help distinguish self from non-self DNA. Future work is necessary to test this hypothesis, but localization could help explain why actively replicating foreign genetic elements are preferentially selected over the chromosome during CRISPR adaptation.

Acknowledgements

B.W. is the founder of SurGene LLC, VIRIS Detection Systems, Inc., and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. Work in the Wiedenheft laboratory is supported by the National Institutes of Health (1R35GM134867 to Blake Wiedenheft), an Amgen Young Investigator award, the M.J. Murdock Charitable Trust, and the Montana State University Agricultural Experimental Station.

References

1 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

2 Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. *J Bacteriol* **194**, 5728-5738, doi:10.1128/JB.01184-12 (2012).

3 Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. *Mol Microbiol* **91**, 900-917, doi:10.1111/mmi.12503 (2014).

4 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:10.1126/science.aal5056 (2017).

5 Klompe, S. E. & Sternberg, S. H. Harnessing "A Billion Years of Experimentation": The Ongoing Exploration and Exploitation of CRISPR-Cas Immune Systems. *CRISPR J* **1**, 141-158, doi:10.1089/crispr.2018.0012 (2018).

6 Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510, doi:10.1038/nature14302 (2015).

7 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res* **40**, 5569-5576, doi:10.1093/nar/gks216 (2012).

8 Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42**, 2483-2492, doi:10.1093/nar/gkt1154 (2014).

9 Liu, T. *et al.* Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. *Nucleic Acids Res* **45**, 8978-8992, doi:10.1093/nar/gkx612 (2017).

10 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202, doi:10.1038/nature14245 (2015).

11 Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-361, doi:10.1101/gad.257550.114 (2015).

12 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740, doi:10.1099/mic.0.023960-0 (2009).

13 Vorontsova, D. *et al.* Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res* **43**, 10848-10860, doi:10.1093/nar/gkv1261 (2015).

14 Richter, C., Gristwood, T., Clulow, J. S. & Fineran, P. C. In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype IF CRISPR/Cas System. *PloS one* **7**, e49549 (2012).

15 Rollins, M. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci U S A* **114**, E5113-E5121, doi:10.1073/pnas.1616395114 (2017).

16 Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci U S A* **114**, E5122-E5128, doi:10.1073/pnas.1618421114 (2017).

17 Almendros, C., Guzman, N. M., Garcia-Martinez, J. & Mojica, F. J. Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. *Nat Microbiol* **1**, 16081, doi:10.1038/nmicrobiol.2016.81 (2016).

18 Datsenko, K. A. & Wanner, B. L. One-Step Inactivation of Chromosomal Genes in Escherichia coli K-12 Using PCR Products. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6640-6645 (2000).

19 Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res* **42**, 8516-8526, doi:10.1093/nar/gku527 (2014).

20 Staals, R. H. *et al.* Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun* **7**, 12853, doi:10.1038/ncomms12853 (2016).

21 Rollins, M. F. *et al.* Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell* **74**, 132-142 e135, doi:10.1016/j.molcel.2019.02.001 (2019).

22 McKenzie, R. E., Almendros, C., Vink, J. N. A. & Brouns, S. J. J. Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nat Protoc* **14**, 976-990, doi:10.1038/s41596-018-0123-5 (2019).

23 Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607-2608, doi:10.1093/bioinformatics/btp450 (2009).

24 Pagès H, A. P., Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. *R package version 2.52.0* (2019).

25 Wickham, H. ggpubr: 'ggplot2' Based Publication Ready Plots. *R package version 0.2.3* (2019).

26 Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *R package version 3.2.1* (2019).

27 Jackson, S. A., Birkholz, N., Malone, L. M. & Fineran, P. C. Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. *Cell Host Microbe* **25**, 250-260 e254, doi:10.1016/j.chom.2018.12.014 (2019).

28 Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res* **42**, 5907-5916, doi:10.1093/nar/gku226 (2014).

29 Lilly, J. & Camps, M. Mechanisms of Theta Plasmid Replication. *Microbiol Spectr* **3**, PLAS-0029-2014, doi:10.1128/microbiolspec.PLAS-0029-2014 (2015).

30 Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev* **72**, 642-671, Table of Contents, doi:10.1128/MMBR.00020-08 (2008).

31 Cheng, K. C. & Smith, G. R. Recombinational hotspot activity of Chi-like sequences. *J Mol Biol* **180**, 371-377, doi:10.1016/s0022-2836(84)80009-1 (1984).

32 Cheng, K. C. & Smith, G. R. Cutting of chi-like sequences by the RecBCD enzyme of Escherichia coli. *J Mol Biol* **194**, 747-750, doi:10.1016/0022-2836(87)90252-x (1987).

33 Rozkov, A. *et al.* Characterization of the metabolic burden on Escherichia coli DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. *Biotechnol Bioeng* **88**, 909-915, doi:10.1002/bit.20327 (2004).

34 Ow, D. S. W., Lee, D. Y., Tung, H. H. & Lin-Chao, S. Plasmid Regulation and Systems-Level Effects on Escherichia coli Metabolism. *Systems Biology and Biotechnology of Escherichia Coli*, 273-294, doi:10.1007/978-1-4020-9394-4_14 (2009).

35 Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141-145, doi:10.1038/s41586-020-2018-1 (2020).

36 Budhathoki, J. B. *et al.* Real-time observation of CRISPR spacer acquisition by Cas1-Cas2 integrase. *Nat Struct Mol Biol* **27**, 489-499, doi:10.1038/s41594-020-0415-7 (2020).

37 Garrett, S. *et al.* Primed CRISPR DNA uptake in Pyrococcus furiosus. *Nucleic Acids Res* **48**, 6120-6135, doi:10.1093/nar/gkaa381 (2020).

38 Tang, J., Akerboom, J., Vaziri, A., Looger, L. L. & Shank, C. V. Near-isotropic 3D optical nanoscopy with photon-limited chromophores. *Proc Natl Acad Sci U S A* **107**, 10068-10073, doi:10.1073/pnas.1004899107 (2010).

39 Yao, S., Helinski, D. R. & Toukdarian, A. Localization of the naturally occurring plasmid ColE1 at the cell pole. *J Bacteriol* **189**, 1946-1953, doi:10.1128/JB.01451-06 (2007).

40 Edgar, R. *et al.* Bacteriophage infection is targeted to cellular poles. *Mol Microbiol* **68**, 1107-1116, doi:10.1111/j.1365-2958.2008.06205.x (2008).

41 Joyeux, M. Preferential Localization of the Bacterial Nucleoid. *Microorganisms* **7**, doi:10.3390/microorganisms7070204 (2019).

Supplemental Figures

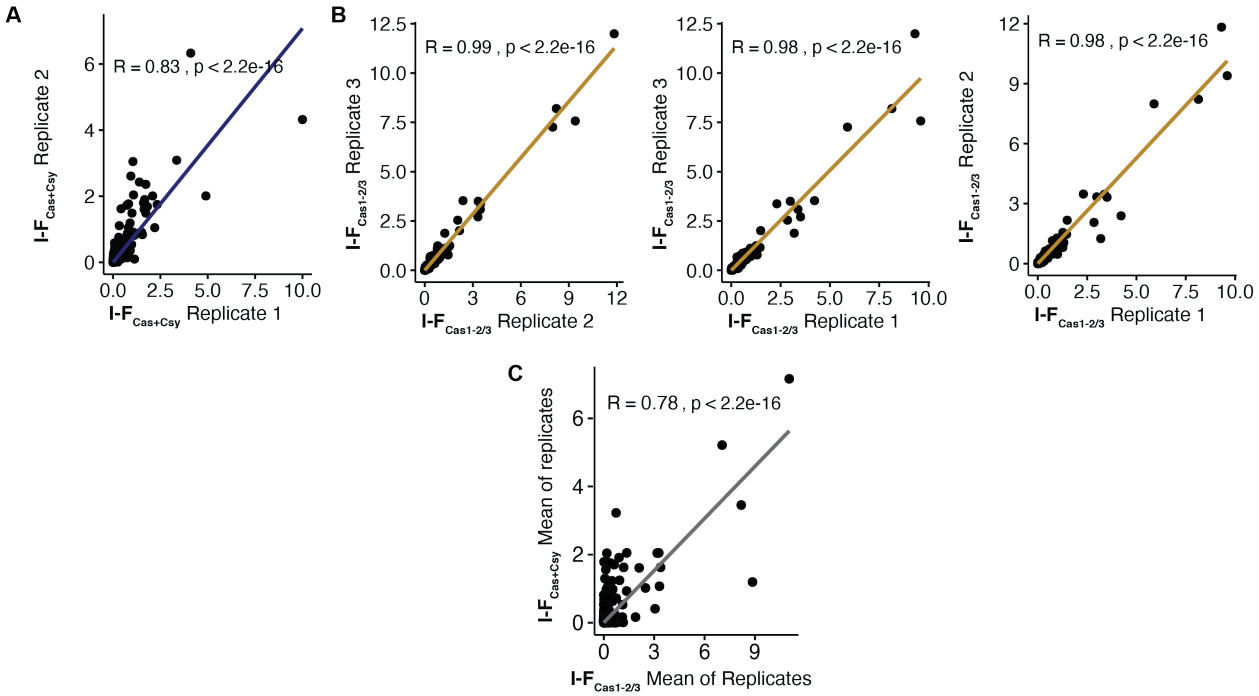


Figure S1.1. Pearson correlation analysis of uniquely aligned pCas1-2/3 protospacer quantities and locations. (A, B) Percentage of total spacers at each position of pCas1-2/3 was used to calculate Pearson coefficients for each biological replicate against other replicates in the same condition. (C) Means of total spacer percentages for two I-$F_{Cas+Csy}$ biological replicates were used to calculate Pearson coefficients against three means of total spacer percentages for I-$F_{Cas1-2/3}$ biological replicates.

Supplemental Tables

Primer	Sequence	Description
P1	5'-GGACGATTTCTTACGCCCTTAT-3'	Forward "internal"
P2	5'-GATGGCATCGATCACACTCAGC-3'	Reverse "internal" and "degenerate"
P3	5'-GCCGTGTAGGCAGCTAAGAAAB-3'	Forward "degenerate"
P-HTS-F1	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTGTGCCGTGTAGGCAGCTAAGAAAB-3'	Forward HTS of I-$F_{Cas1-2/3}$ replicate 1
P-HTS-F2	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCTGCCGTGTAGGCAGCTAAGAAAB-3'	Forward HTS of I-$F_{Cas1-2/3}$ replicate 2
P-HTS-F3	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCAGAGCCGTGTAGGCAGCTAAGAAAB-3'	Forward HTS of I-$F_{Cas1-2/3}$ replicate 3
P-HTS-F4	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGCCTGCCGTGTAGGCAGCTAAGAAAB-3'	Forward HTS of I-$F_{Cas+Csy}$ replicate 1
P-HTS-F5	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGGAGCCGTGTAGGCAGCTAAGAAAB-3'	Forward HTS of I-$F_{Cas+Csy}$ replicate 2
P-HTS-R	5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATGGCATCGATCACACTCAGC-3'	Reverse HTS for all samples
P-Csy_{PAMmut}-F	5'-CAAGCCGCAGGCGATCAGTCAGTTGAAC-3'	Q5 mutagenesis of pCsy
P-Csy_{PAMmut}-R	5'-GGTCTTCAAGCCGCCATG-3'	Q5 mutagenesis of pCsy

Table S1.1. Primers used in these experiments.

					New Spacers				Aligned Spacers (31-33bp)*		
	Condition	Biological Replicate	Total HTS Reads	Spacers Extracted	Total	31bp	32bp	33bp	Genome	pCas1-2/3	pCsy
BC1	pCas1-2/3 Only	#1	2,485,942	368,477	325,852	7,181	299,831	6,119	99	172,446	
BC2		#2	3,203,584	734,524	698,002	10,922	651,127	10,226	128	366,879	
BC3		#3	2,614,507	636,992	603,640	11,436	561,265	10,082	84	283,893	
BC9	pCsy + pCas1-2/3	#1	6,309,887	2,604,598	2,604,510	32,801	2,418,309	35,251	105	1,706,595	1,424,206
BC10		#2	3,511,329	1,440,015	1,439,981	14,184	1,345,942	17,756	170	904,611	738,699

Table S1.2. Summary of High-throughput sequencing data.

CHAPTER TWO

DISTRIBUTION AND PHASING OF SEQUENCE MOTIFS THAT FACILITATE CRISPR ADAPTATION

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Andrew Santiago-Frangos

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript

Co-Author: Murat Buyukyoruk

Contributions: designed the research, prepared manuscript figures, performed the research

Co-Author: Tanner Wiegand

Contributions: designed the research, prepared manuscript figures, performed the research

Co-Author: Pushya Krishna

Contributions: designed the research, performed the research

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

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

Current Biology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Elsevier, Inc.
Volume 31
Issue 16
23 August 2021
Pages 3515-3524
DOI: 10.1016/j.cub.2021.05.068

DISTRIBUTION AND PHASING OF SEQUENCE MOTIFS THAT FACILITATE CRISPR ADAPTATION

Andrew Santiago-Frangos,[1] Murat Buyukyoruk,[1,2] Tanner Wiegand,[1,2] Pushya Krishna,[1,2] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]These authors contributed equally

Abstract

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

Introduction

The repeat-spacer-repeat architecture characteristic of all CRISPR (clustered regularly interspaced short palindromic repeat) loci was first observed in 1987.[1] However, the biological function of CRISPRs remained obscure until 2005, when three groups independently reported that CRISPR loci frequently contain "spacers" derived from foreign genetic elements.[2-4] By comparing CRISPR loci from closely related isolates of *Yersinia pestis*, Pourcel et al.[4] noted that new phage-derived spacers are preferentially added to one end of the CRISPR, which is often flanked by an adenine and thymine-rich (AT-rich) sequence, previously termed the "leader."[5] Collectively, these computational observations indicated that CRISPRs are part of an adaptive immune system that maintains a chronological record of previously encountered foreign genetic parasites.[6,7] A role for CRISPR loci and Cas proteins in adaptive immunity was first established by phage challenge experiments performed by Barrangou et al.[8] Polarized integration is crucial in most systems, because spacers at the leader end of the CRISPR have been shown to provide greater levels of immunity.[9]

Structural and biochemical experiments aimed at understanding the mechanism(s) of polarized adaptation have resulted in two models for preferential integration of new spacer at the "leader end." In type II systems, the Cas1 proteins have been shown to recognize the leader-repeat junction, while type I-E and I-F systems have been shown to rely on a DNA-bending protein called integration host factor (IHF).[9-15] In the type I-E CRISPR system of *Escherichia coli* (K12), IHF binds to a roughly 30-bp sequence motif that begins 6 bp upstream of the first repeat.[10] IHF binding kinks DNA in the leader, creating a horseshoe-shaped structure that stabilizes the Cas1-2 integrase complex bound to the first repeat of the CRISPR locus.[11] One of

the Cas1 dimers is wedged into the "toe" of the DNA horseshoe, resulting in specific contacts between the Cas1 and IHF proteins, as well as sequence-specific interactions between one lobe of Cas1 and an upstream sequence motif.[10,11]

We recently noticed that the IHF binding site in the leader sequence of the type I-F CRISPR locus from *Pseudomonas aeruginosa* (PA14) is 8 bp further from the leader-repeat junction than what was originally observed in the type I-E systems of *Escherichia coli* (Figure 2.1). We hypothesized that this seemingly subtle difference has important mechanistic implications for CRISPR adaptation. The helical structure of double-stranded DNA (dsDNA) contains ~11 bp per 360-degree rotation.[16,17] Thus, the addition of 8 bp not only shifts the IHF binding site ~27 Å away from the leader-repeat junction, which would preclude previously observed Cas1-IHF interactions, but this insertion also introduces an ~260-degree rotation of the upstream DNA. This rotation would position one tip of the "DNA horseshoe" on the opposite side of the Cas1-2 integration complex from what has been previously observed for the *E. coli* integration complex.[11]

Here, we analyze 14,095 bacterial and 1,179 archaeal CRISPR leaders for conserved sequence motifs. This analysis reveals discretely distributed IHF binding sites and upstream motifs in a subset of I-C, I-E, I-F, and II-C CRISPR loci. These leaders frequently contain multiple IHF binding sites and subtype-specific upstream motifs. The sequence, spacing, and orientation between motifs vary within and between subtypes, resulting in the detection of approximately 20 distinct leader architectures (Figure S2.1). Differences in leaders between closely related strains frequently involve insertions or deletions (indels) of 10–12 bp, which preserves the phase of these motifs, suggesting that phase is more important than distance per se.

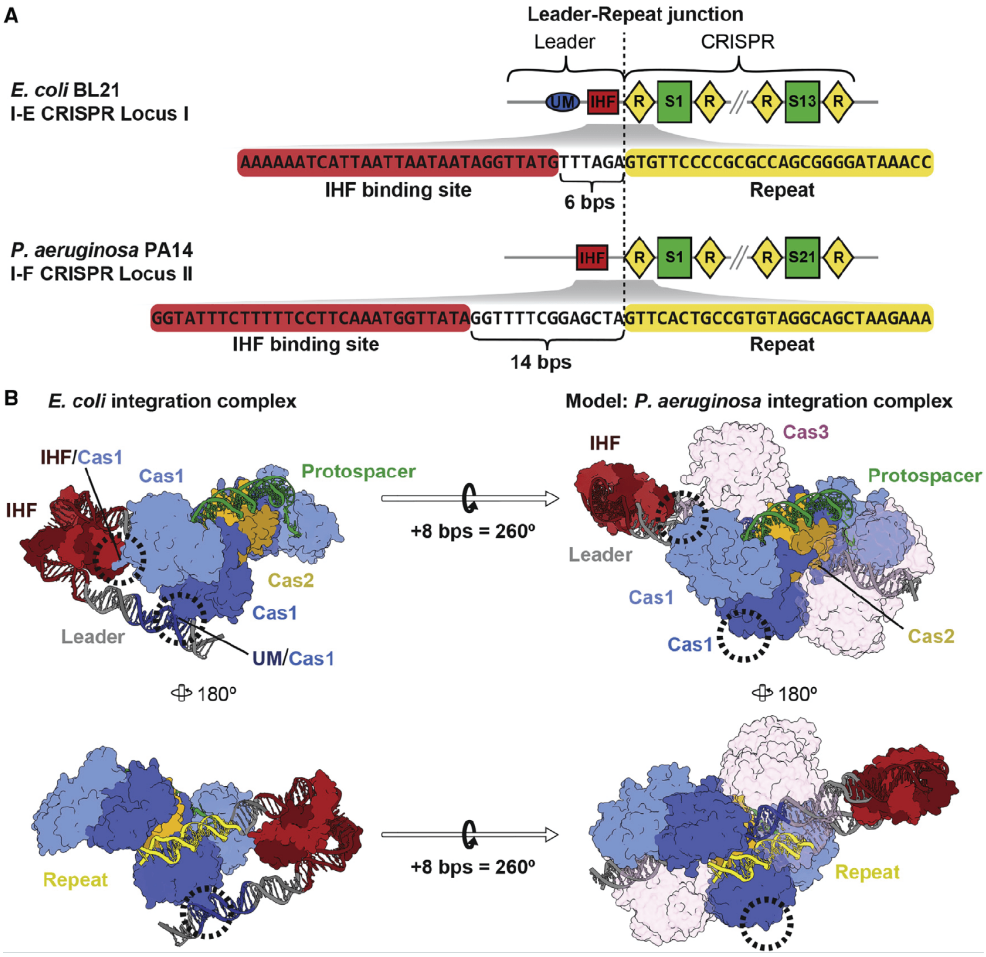


Figure 2.1. Spacing of sequence motifs in CRISPR leaders has mechanistic implications for the mechanism of spacer integration. (A) Schemes of two CRISPR loci from *E. coli* BL21 (type I-E, top) and *P. aeruginosa* PA14 (type I-F, bottom) and sequences of the first CRISPR repeat and the upstream IHF-binding site in the leaders. Repeating DNA motifs (diamonds) in the CRISPR locus are interspersed with unique spacer sequences (rectangles). The IHF-binding site in the I-F system from *P. aeruginosa* PA14 is 8 bp further away from the leader-repeat junction, as compared to the I-E CRISPR locus. (B) Structure of the previously determined type I-E integration intermediate (PDB: 5WFE). The Cas1-2 hetero-hexamer (blue and yellow) is in the process of integrating a protospacer (green) at the first CRISPR repeat (yellow). IHF kinking of DNA at the IHF-binding site recruits an upstream motif (UM) (dark blue) to interact with one lobe of a Cas1 subunit, while the second Cas1 subunit in the same dimer makes protein-protein contacts with IHF.11 Models of the 8-bp insertion present in *P. aeruginosa* CRISPR locus II illustrate how this insertion is predicted to disrupt both of these interactions by shifting IHF 27 Å away from the integration complex and rotating the UM 260º. Cas3 domains of the I-F Cas2/3 fusion are shown semi-transparently for clarity. See also Figures S1 and S3.

Additionally, we use *in vitro* adaptation assays to test the importance of both the sequence and the phase of leader motifs on new spacer integration in the type I-F system from *P. aeruginosa*. Overall, our data suggest that the mechanisms of polarized CRISPR adaptation are diverse but all systems that rely on IHF are expected to be phase dependent.

Results

Identification and distribution of IHF and upstream motifs

IHF binding sites have been identified in the leader sequences of a few type I-E systems and in the type I-F system from *Pectobacterium atrosepticum*.[10,12,18] To broadly determine the prevalence and distribution of IHF binding sites in CRISPR leaders, we queried all complete bacterial and archaeal genome sequences available at NCBI for CRISPR loci and associated leader sequences using CRISPRDetect.[19] In total, we identified 15,274 CRISPR loci, representing most of the major subtypes (Table 1). We queried 200 bp upstream of each CRISPR using previously established position weight matrices for IHF binding sites.[20,21] Sequences identified with this position weight matrix were used to construct more detailed position weight matrices, which were then used to iterate the search. In total, we identified 8,631 putative IHF binding sites in 15,274 leader sequences. Next, we culled the list of leaders by eliminating redundant sequences that were greater than 95% identical. This eliminates closely related genomes from a few organisms that were oversampled in our initial dataset (e.g., *E. coli*) and enables a more accurate representation of the phylogenetic distribution of IHF binding sites in leader sequences. In this subset of 6,533 non-redundant leaders, IHF binding sites were found in 24% of I-C (n=176), 26% of I-E (n=336), 83% of I-F (n= 444), and 32% of II-C (n = 108) leaders (Table 1).

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

Table 2.1. Subtype distribution of microbial CRISPRs and identified IHF-binding sites. 15,274 CRISPR loci were identified in bacterial and archaeal genomes. Closely related microbial genomes, which cause an over-representation of certain CRISPR subtypes (I-B, I-C, I-E, I-F, II-A, II-C, and III-A; "redundant" column), were systematically removed ("non-redundant" columns). The number of non-redundant CRISPR leaders in containing at least one or at least two unique IHF-binding sites is reported. See also Figure S2.1.

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

According to the previously established model for type I adaptation, the DNA-bending behavior of IHF presents an upstream sequence motif that is recognized by specific amino acids on Cas1 (i.e., R131 and R132).[11] In support of this model, the proximal IHF binding sites in I-E leaders are tightly distributed around a midpoint of 20 bp from the leader-repeat junction, and I-C and II-C leaders possess similar IHF sites at midpoints of 22 and 21 bp from the leader-repeat junction (Figure S2.1). However, the additional 8 bp that separate the IHF binding site from the leader-repeat junction in *P. aeruginosa* PA14, as compared to *E. coli* BL21, are preserved in a comparison of all type I-E and I-F leaders that contain IHF binding sites (Figure S2.1). This ~8-bp insertion is unique to the type I-F systems and may be an evolutionary adaptation necessary to accommodate the Cas3 domain, which is uniquely fused to the Cas2 protein in I-F CRISPR system (Figure 2.1).[22-24] In addition to making room for Cas3, this insertion also rotates the upstream DNA by ~260 degrees. This rotation suggests that either the Cas1-2/3 integration complex relies on a distinct mechanism for recognizing an upstream sequence motif or that I-F systems do not rely on upstream sequence motifs. To distinguish between these two possibilities and to identify conserved upstream sequence motifs in other IHF-containing leaders, we performed a *de novo* motif search using a combination of MEME and FIMO.[21,25] MEME identified the upstream motif previously reported for type I-E systems,[11,18] as well as several new motifs (Figure 2.2). The sequence and location of these motifs are characteristic of specific subtypes. In addition to the upstream motifs and proximal IHF binding sites, MEME also identified additional upstream IHF-binding sites and additional, distally located upstream motifs that are either direct repeats or inverted repeats of the proximal upstream motif (Figure 2.2). The spacing between these motifs sometimes varies between leaders, but changes in the spacing are

almost always restricted to increments of 10– 12 bp, indicating that phase is conserved (Figure S2.1). The I-C distal direct repeat and inverted repeat of the upstream motif are exceptions to this rule, which are often shifted by 8 bp (Figure S2.1).

While IHF has been shown to be critical for polarized CRISPR adaptation in the type I-E system from *E. coli*,[10,11] we found that roughly 75% of I-E leaders do not contain a canonical IHF binding site. 20% of I-E leaders instead contain a "leader-anchoring motif" that directly abuts the first repeat, in a similar manner to the "leader-anchoring site" reported for several of the type II systems (Figure 2.2).[9,13-15] While proximity of I-E leader-anchoring motifs to the leader-repeat junction is similar to those in type II systems, the sequences themselves are notably different. The type I-E leader-anchoring motifs contain an off-center TCA sequence and a 5' to 3' TTR triplet on the opposite strand, which are hallmarks of IHF and Hbb DNA-binding sites (Figure S2.2),[26-30] though these motifs are not detectable using any of the previously established position weight matrices for IHF binding sites (Figure 2.2). The IHF-like leader-anchoring motif may associate with DNA-bending proteins that recruit upstream motifs. In fact, MEME identifies two motifs (I-E_A and I-E_B) that are unique to I-E leaders that contain "IHF-like leader-anchoring motifs" (Figure 2.2). If the IHF-like leader-anchoring motif is in fact a DNA-bending sequence that functions to recruit the unique upstream motifs (i.e., I-E_A or I-E_B), then the proximity of this motif relative to the leader-repeat junction would change the phase of the upstream motif relative to what has been observed for the IHF-containing leader from E. coli. Thus, we hypothesized that Cas1 proteins associated with these systems may not rely on arginines 131 and 132 (R131 and R132), which are critical for upstream motif recognition by Cas1 in IHF-dependent CRISPRs from *E. coli*.[11] To test this hypothesis, we aligned 368 Cas1

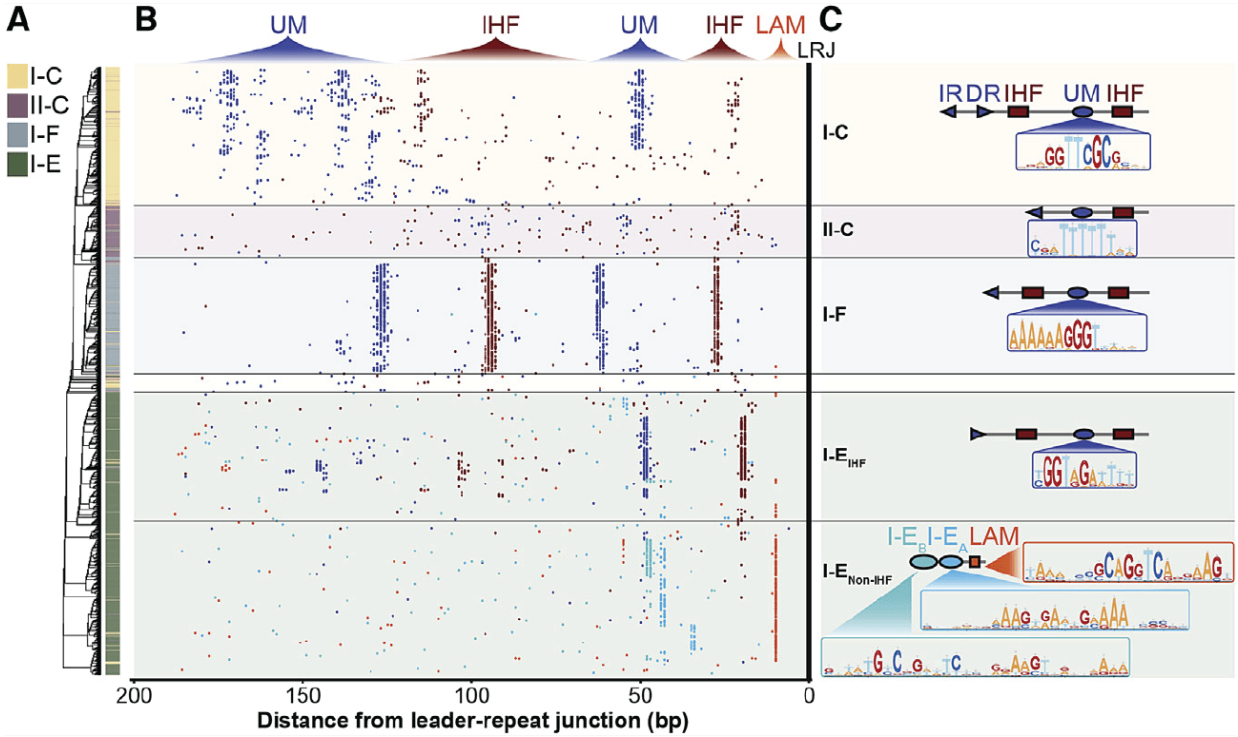


Figure 2.2. IHF-directed CRISPR adaptation is widespread. (A) Phylogenetic tree generated from alignment of 200 bp of I-C, II-C, I-F, and I-E CRISPR leaders and the first repeat. (B) Distributions of motifs within I-C, II-C, I-F, and I- E CRISPR leaders. Each dot represents the midpoint of IHF-binding sites (red), subtype-specific UMs (dark blue), I-E_A (cyan), I-E_B (teal), or leader-anchoring motif (orange). Many of the leaders shown possess proximal IHF and UMs found between 0 and 70 bp upstream of the leader-repeat junction (LRJ) and distal IHF and UMs found 70–200 bp upstream of the LRJ. (C) Schematic of the prominent architecture for motifs within I-C, II-C, I-F, and I-E CRISPR leaders. Position weight matrices are shown for subtype-specific upstream motifs and motifs found in I-E leaders that do not contain IHF- binding sites. See also Figure S2.1.

sequences from CRISPR systems that contain IHF-like leader-anchoring motifs. Unlike I-E Cas1 proteins that are associated with IHF-containing leaders, these Cas1 proteins do not maintain R131 and R132, suggesting that the mechanism of adaptation in these CRISPRs is distinct (Table 2). These findings suggest these two types of I-E CRISPR systems may be phylogenetically distinct; indeed, IHF-regulated I-E CRISPRs are found in bacteria belonging to the Phylum *Proteobacteria*, while leader-anchoring motif-regulated I-E CRISPRs are found in *Actinobacteria*. Collectively, these results reveal diverse upstream architectures that

			R131		R132	
Subtype	Leader motifs	Numbers of Cas1s	Identity	Similarity	Identity	Similarity
I-E	IHF or UM	454	57.5%	68%	71.6%	72.9%
I-E	LAM, I-E_A, I-E_B	368	3.7%	5.4%	11.2%	13.8%
I-F	IHF or UM	499	11%	11%	11.9%	22.4%

Table 2.2. Cas1 residues that interact with canonical I-E UM are only conserved in Cas1s associated with IHF-regulated I-E CRISPRs. See also Figure S2.1.

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

IHF-binding sites and upstream motif are critical for efficient integration

To determine whether the IHF-binding sites and upstream motifs play a direct role in I-F CRISPR adaptation, we performed in vitro integration assays using the leader sequence derived from the CRISPR2 locus of P. aeruginosa PA14. According to our bioinformatic analysis, this leader contains both proximal and distal IHF binding sites, as well as proximal and distal upstream motifs (Figures 2 and 3). To determine which of these motifs participate in new spacer integration, we compared integration efficiencies measured for the wild-type leader sequence to leaders where we either replaced the proximal IHF binding site with an IHF consensus sequence from *E. coli* (Opt. IHF prox), mutated key positions in the proximal IHF binding site (Mut. IHF prox), scrambled the proximal upstream motif (Mut. UM prox), deleted the distal IHF site (Del. IHF distal), or scrambled the distal upstream motif (Mut. UM distal; Figure 2.3C).

IHF-binding sites in the *P. aeruginosa* CRISPR2 leader deviate from the *E. coli* IHF consensus sequence at the 50 “A-tract” and central “WATCAR” regions, despite *P. aeruginosa* and *E. coli* IHF proteins sharing greater than 70% sequence identity (Figure S2.2).[26-28] To

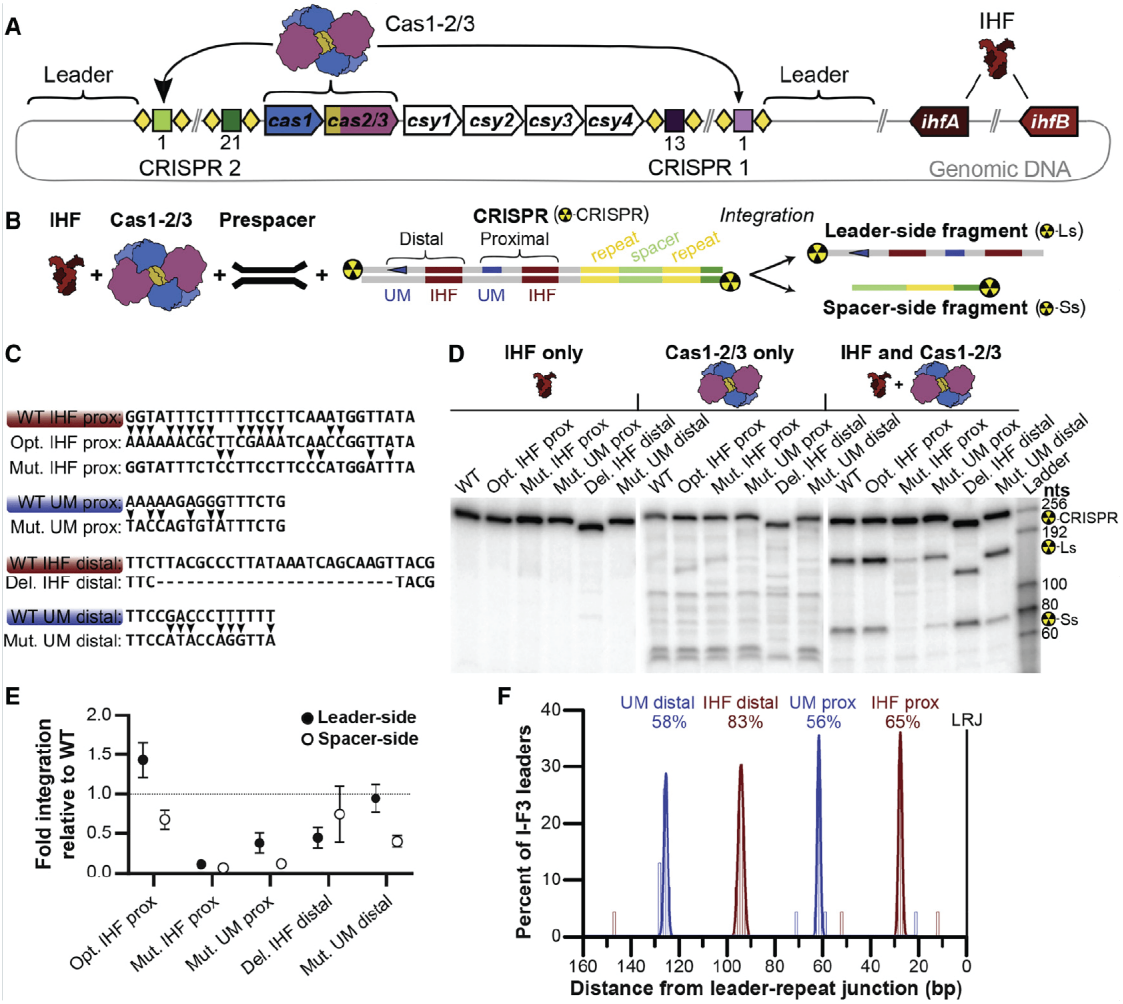


Figure 2.3. IHF-binding sites and UM are critical for efficient integration. (A) Scheme of I-F CRISPR system from *P. aeruginosa* PA14. Two CRISPR loci composed of repeats (diamonds) and spacers flank six *cas* genes (arrows). Cas1-2/ 3 assembles into a heterohexameric complex (blue, yellow, and purple, respectively) that catalyzes the integration of new spacers into CRISPR loci. IHF is a heterodimer of two related proteins encoded by *ihfA* (brown) and *ihfB* (tan). (B) Scheme for *in vitro* integration of a prespacer DNA into 32P-labeled DNA derived from the CRISPR2 locus of *P. aeruginosa* PA14. Leader- and spacer-side transesterification reactions produce a large and a small 32P-labeled DNA fragment, respectively, which are separated by denaturing gel electrophoresis. Proximal and distal IHF-binding sites and UMs in the leader are annotated. (C) IHF-binding sites (brown) or UMs (blue) are shown. Two variants (arrows) of the proximal IHF site were tested (i.e., Opt IHF prox and Mut IHF prox.), and one variant of the proximal UM (Mut. UM prox.), one variant of the distal IHF (Del. IHF distal), and one variant of the distal UM (Mut. UM distal) were tested. (D) Representative images of integration assays resolved on denaturing gels. Reactions were performed either in the presence of IHF alone (left), Cas1-2/3 alone (middle), or in the presence of both protein complexes (right). The expected positions of full-length CRISPR substrate (Del. IHF distal variant, 233 nt; wild

type [WT] and other variants, 255 nt) as well as leader side (Ls) (Del. IHF distal variant, 127 nt; WT and other variants, 151 nt) and spacer side (Ss) (all variants, 77 nt) integration products are indicated. (E) Quantification of leader- (dark gray) or spacer-side (light gray) integration. The average (±SD) of triplicate reactions is shown. Integration was confirmed via high-throughput sequencing (Figure S2.3). Uncropped images and replicate gels are provided (Figure S2.4). (F) Tally of the midpoints of IHF-binding site (red) and I-F upstream motifs (blue) identified in 23 non-redundant I-F3 CRISPR leaders.31 The area of histograms fit to peaks of motif matches is reported to convey the percentage of leaders with a given motif. Most I-F3 (Tn7-associated) CRISPR leaders possess motifs needed for efficient spacer integration. See also Figures S3–S5.

determine whether an "optimized" IHF site would result in more efficient spacer integration, we replaced the IHF binding site in the I-F leader with the consensus IHF-binding sequence from *E. coli*. These two sequences differ at 15 positions over 29 bp. Although optimization of the IHF-binding site (*Opt. IHF prox*) resulted in slightly increased leader-side integration (1.4-fold), there is a concomitant decrease in spacer-side integration (1.5-fold; Figures 3D and 3E). To verify that the IHF-binding sequence is important for integration, we mutated six bases that are critical for IHF recognition (Figure 2.3C).[26,27,31] These mutations (*Mut. IHF prox*) result in large decreases for both leader and spacer-side integration (8.7and 14.7-fold, respectively; Figures 3D, 3E, S3, and S4), which supports previous work from the Fineran lab demonstrating that the IHF protein is necessary for adaptation in the I-F system from *P. atrosepticum*.[12] While it is expected that IHF is involved in DNA bending, there have been no prior reports of an upstream motif in I-F CRISPR leaders. To determine whether the proximal upstream motif sequence we identified is important for CRISPR adaptation, we scrambled the sequence (*Mut. UM prox*). This mutation results in a large decrease in both leader and spacer-side integration efficiency (2.6 and 8.3-fold, respectively; Figures 3C–3E, S3, and S4).

Fagerlund et al.[12] previously detected several IHF-like sequences upstream of the I-F CRISPR locus in *P. atrosepticum*, but the importance of these distal sequences has gone

untested. We hypothesized the distal IHF-binding site and the distal upstream motif, which both occur at high frequency, would be functionally important for CRISPR adaptation. To test this hypothesis, we deleted the distal IHF site (*Del. IHF distal*) or scrambled the distal upstream motif (*Mut. UM distal*). Deletion of the distal IHF site reduces leader-side integration by 2.2-fold while changes to spacer-side integration are within experimental error (Figures 2D and 2E). In contrast, scrambling the distal upstream motif sequence does not impact leader-side integration but reduces spacer-side integration by 2.5-fold. Collectively, these results suggest that the distal IHF and upstream motif sequences participate in new spacer integration.

Tn7-associated I-F3 leaders contain motifs needed for efficient integration

Recently, Petassi et al.[32] reported that Tn7-associated I-F3 CRISPR systems are accompanied by short CRISPRs. The spacers in I-F3 CRISPRs have recently been shown to guide transposition in a sequence-specific manner.[33,34] To determine whether I-F3 CRISPR loci may continue to acquire new spacers that would enable the Tn7-like system to adapt to transpose into novel locations, we analyzed a subset of I-F3 leaders using custom position weight matrices for IHF-binding sites and I-F upstream motifs developed above. In 23 non-redundant I-F3 leaders, 65% and 85% possess proximal and distal IHF-binding sites (respectively) and 56%–58% possess proximal and distal IF upstream motifs (Figure 2.3F). Further, all four leader motifs are found in the same positions as in canonical I-F leaders (Figure 2.2). Because most I-F3 systems are not associated with operons that encode *cas1* or *cas2/3*, it has been hypothesized the integration complex may be provided in *trans*.[35] The conservation of IHF-binding sites and upstream motifs in a subset of the I-F3 leaders suggests that *trans*-acting Cas1-2/3 integration complexes may be compatible with these leaders. One barrier to the spread of mobile genetic

elements is compatibility with host factors. Our *in vitro* integration results show that *P. aeruginosa* IHF can recognize an *E. coli* consensus IHF-binding site (Figures 3C–3E), corroborating previous reports that *E. coli* IHF may complement the integration of new spacers into I-F CRISPR loci by Cas1-2/ 3.[36,37] Further, these observations suggest that I-C, II-C, I-E, I-F, and Tn7-associated I-F3 CRISPR systems may exchange efficiently between microbes that encode IHF.

Phased leader motifs facilitate CRISPR adaptation

The insertion of 1, 2, or 5 bp between the leader-repeat junction and the IHF-binding site was previously shown to progressively inhibit new spacer integration into a I-E CRISPR locus.[11] These results were used to highlight the importance of the IHF-Cas1 interaction, which we have no reason to question. However, we hypothesized that these mutations would also alter the position of the upstream motif and that phase of the upstream motif might be more important than distance per se. To test this hypothesis, we measured the integration efficiency of new spacers into a fragment of a CRISPR from strain 14 of *P. aeruginosa* (PA14). Variants of the PA14 leader were designed to either preserve (e.g., +10 bp) or disrupt (e.g., +5 bp) phasing between either the leader-repeat junction and the proximal IHF, the proximal IHF site and the proximal upstream motif, or both (Figure 2.4A).

In the absence of IHF, the Cas1-2/3 proteins from *P. aeruginosa* inefficiently integrate prespacers into multiple sites along the leader (Figures 3D and 4B). In contrast, the addition of IHF to these reactions facilitates Cas1-2/3-mediated integration of prespacers at the leader and spacer side of the first repeat (Figures 3D and 4B). High-throughput sequencing of the integration reactions performed using primers at either end of the CRISPR DNA confirm that

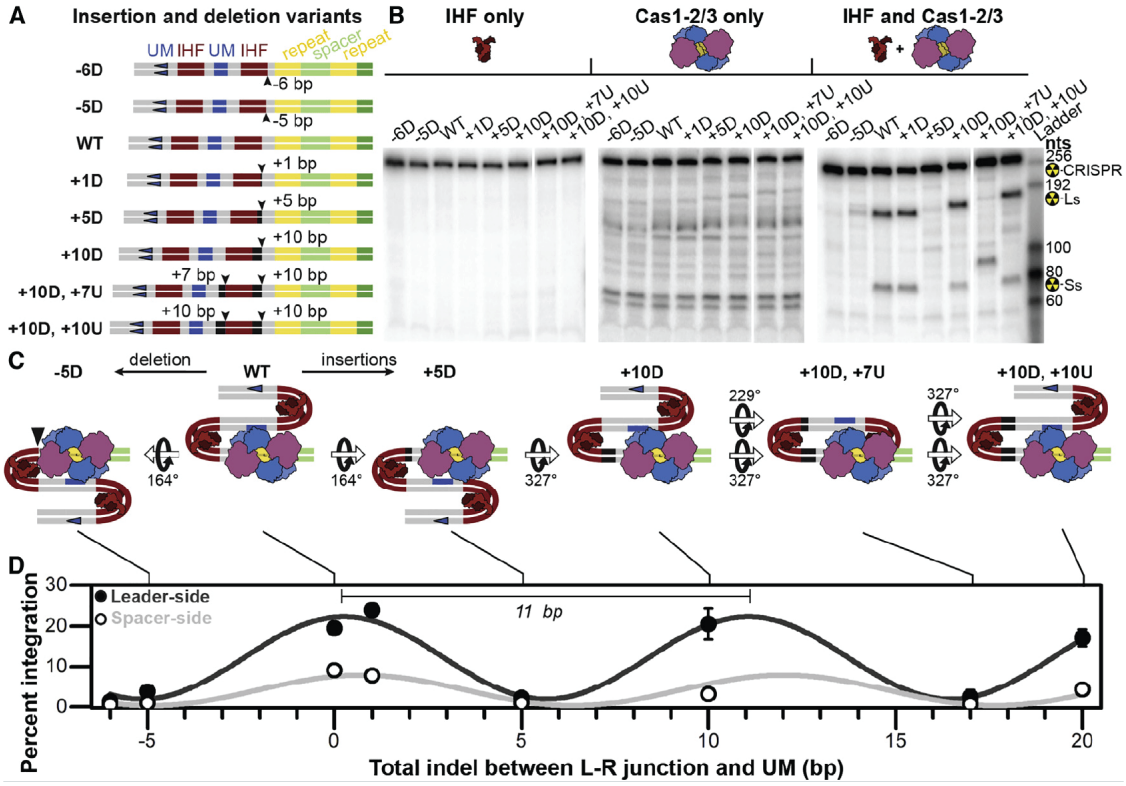


Figure 2.4. Phase-dependent CRISPR adaptation. (A) Schematic of insertion and deletion (indel) variants generated to test the impact of distance and phase of leader motifs relative to the CRISPR locus. (B) I-F CRISPR integration is insensitive to indels that maintain phase (10 bp), while indels of less than a full helical turn (5–7 bp) result in integration defects. Representative images of integration assays resolved on denaturing gels are shown. Reactions were performed either in the presence of IHF alone (left), Cas1-2/3 alone (middle), or in the presence of both protein complexes (right). The expected positions of full-length substrate (249–275 nt, depending on indel) as well as leader-side (145–161 nt, depending on indel) and spacer-side (77 nt for all variants) integration products are indicated. (C) Schemes of putative IHF-bound conformations of WT and mutant leaders. Estimated rotations of DNA, relative to WT, are indicated. (D) Quantification of prespacer integration into all leader indel variants, either at leader (dark gray) or spacer side (light gray). The average (±SD) of triplicate reactions is shown. Leader- and spacer-side integration data were independently fit to sine waves (R2 = 0.95 and R2 = 0.75) of wavelengths 10.8 ± 0.4 and 11.3 ± 0.7 bp, respectively. Integration was confirmed via high-throughput sequencing (Figure S2.3). Uncropped images and replicate gels are provided (Figure S2.6). See also Figures S2.3 and S2.6.

IHF decreases off-target spacer integration from 75.2% to 9.0% (Figure S2.3). Additionally, we observed that leader-side integration is more efficient than spacer-side integration and that 49.8% of unambiguously integrated prespacer substrates are trimmed prior to integration (Figure S2.3;

see STAR Methods for explanation of ambiguous spacer integration events). These results are similar to those previously reported for *E. coli* I-E and *P. atrosepticum* I-F systems.[10,12,38]

To specifically test the importance of phase, we designed a series of I-F leader mutants that maintain the motif but perturb the phase of the proximal IHF and proximal upstream motifs to varying degrees (6D, 5D, +1D, and +5D; Figure 2.4A). Consistent with our hypothesis, indels between the proximal IHF site and the leader-repeat junction, or the proximal IHF site and proximal upstream motif, both inhibit new spacer integration. The degree to which these changes impact the efficiency of integration correlates with changes to the phase. In other words, the more out of phase, the lower the efficiency of integration. Importantly, insertions that restore the original phase (i.e., +10D and +10U) also restore integration activity (Figure 2.4). Collectively, the data show that integration efficiencies oscillate with a wavelength of ~11 bp, corresponding to a complete turn of double-stranded DNA helix (Figure 2.4D). These experiments indicate that the phase of leader motifs, rather than their distance from the leader-repeat junction, is critical for efficient integration of new spacers at the leader end of the CRISPR locus.

Discussion

Since the discovery of regulatory elements in DNA,[39] there has been an interest in determining the "grammatical rules" of sequence motifs that regulate the storage and retrieval of genetic information. The regulatory influence of a sequence motif is determined by its location and orientation to other *cis* or *trans* acting sequence motifs, which often assemble into structures that regulate the activity of neighboring genetic elements.[40-42]

Specific and efficient integration of new spacers into CRISPR loci is the crucial first step in the adaptive immune response of bacteria and archaea. Here, we identify leader sequences and

structural elements that regulate CRISPR adaptation. These grammatical rules help explain the mechanisms by which CRISPRs evolve and may be important for the design of CRISPR-based-data-recording applications.[43,44] Moreover, we anticipate that motifs identified here may be used to improve computational methods designed to identify and assign CRISPR loci to particular subtypes.

While new spacers are generally added to one end of the CRISPR, there are several known examples of non-random integration of spacers outside the CRISPR locus.[10,12,14,45] To determine whether off-target integration observed in *E. coli* could be explained by the presence of specific motifs, we queried 200 bp flanking both sides of 697 integration sites to identify I-E repeats, IHF-binding sites, and upstream motifs using FIMO.[21,45] Only 73 of these sites possess a significant match to a I-E repeat, 31 possess an IHF-like binding site, and 44 possessed a canonical I-E upstream motif. Importantly, none of these sites contain all three motifs. While the mechanistic basis for integration at these non-CRISPR sites remains unclear, the overexpression of Cas1 and Cas2 may explain some of the promiscuity reported in these experiments.[45]

Previous studies have demonstrated that some type I CRISPRs rely on leader proximal and distal motifs.[10-12,18,46-48] Conversely, in type II systems, Cas1-2 directly recognizes a leader-anchoring motif located directly adjacent to the first repeat.[9,13-15,49] Our work suggests that the mechanisms for adaptation may be more complex and that some II-C CRISPRs may require IHF, while 20% of type I-E CRISPRs appear to rely on a leader-anchoring motif that is unique in sequence but similar in position to what has been observed for the type II systems (Figure 2.2). Bioinformatic analyses suggest IHF is predominantly restricted to *Proteobacteria*, although similar type II DNA-binding proteins may be found in other microbes.[29,50] Perhaps as expected,

CRISPR leaders in which we identified phased IHF-binding sites and upstream motifs (I-E, I-F, I-C, and II-C) are highly represented in *Proteobacteria*.51 Further, the split between I-E leaders that contain IHF sites versus those that contain leader-anchoring motifs roughly coincides with a split between I-E CRISPRs originating from *Proteobacteria* versus those originating from *Actinobacteria*. These observations point to a continuing co-evolution of CRISPR adaptation complexes with host machinery to support the fidelity of new spacer integration.

Leader sequences are diverse, and variations in leader architectures may tune adaptation rates. In fact, many microbes possess multiple CRISPR loci, which evolve at different rates.[51-53] For example, I-F *cas* genes in *P. aeruginosa* are flanked by a short CRISPR locus (CRISPR1) that evolves slowly and a longer CRISPR locus (CRISPR2) that has been shown to evolve more rapidly (Figure S2.5).[53,54] *P. atrosepticum* SCRI1043 similarly possesses three CRISPR loci that evolve at different rates.[52] In corroboration with the biochemical data presented here, the fastest evolving I-F CRISPR loci in *P. aeruginosa* PA14 and *P. atrosepticum* SCRI1043 contain leaders with proximal and distal IHF-binding sites, as well as upstream motifs, whereas slower evolving CRISPR loci in these bacteria lack the proximal IHF-binding site in their respective leaders (Figure S2.5). Collectively, these data suggest that CRISPR loci with leaders containing the full complement of motifs may adapt rapidly. Although rapidly evolving CRISPR loci may enable a quick response to a new pathogen, the dsDNA breaks associated with the integration of a new spacer may cause such loci to be prone to lose previously acquired spacers via homologous recombination between repeats or other DNA repair pathways.[55,56] Thus, there may be a selective pressure that maintains CRISPR loci that have been tuned for either fast or slow adaptation.

The work presented here started with the observation that there are 8 additional base pairs between the leader-repeat junction and the first repeat of the CRISPR in *P. aeruginosa* (PA14), as compared to the CRISPRs in *E. coli*. This 8-bp insertion is not unique to the CRISPR in *P. aeruginosa* (PA14) but is instead conserved in I-F systems (Figure S2.1). While we show that IHF-binding sites are common in I-E and I-F systems, they are certainly not the rule. In fact, 75% of non-redundant I-E leaders do not contain IHF-binding sites. Rather, 20% of I-E systems contain a leader-anchoring motif and one or two different unique upstream sequence motifs (Figure 2.2). These observations led us to look more broadly for IHF-binding sites and unique upstream motifs, revealing that many leaders contain multiple IHF-binding sites and diverse upstream sequence motifs that are characteristic of the I-C, I-E, I-F, and II-C leaders. The conservation of relative distances between motifs suggested that phase, rather than distance, might be critical for efficient adaption in IHF-dependent systems. We tested the importance of phase in the I-F system and demonstrate insertions that are in phase with the DNA helix restore efficient adaption (Figure 2.4). While leaders are expected to be critical for adaption in most systems, most leaders are also expected to control CRISPR expression.[57-59] Ultimately, we expect that the motifs identified here are only the tip of the iceberg and that additional bioinformatic studies will provide ongoing insights that will lead to a more comprehensive understanding of how leader sequences integrate signals for expression and adaptation of CRISPRs.

Acknowledgements

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

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

References

1 Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology* **169**, 5429-5433, doi:10.1128/jb.169.12.5429-5433.1987 (1987).

2 Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561, doi:10.1099/mic.0.28048-0 (2005).

3 Mojica, F. J., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution* **60**, 174-182 (2005).

4 Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653-663, doi:10.1099/mic.0.27437-0 (2005).

5 Jansen, R., Embden, J. D. A. v., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43**, 1565-1575, doi:10.1046/j.1365-2958.2002.02839.x (2002).

6 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:10.1126/science.aal5056 (2017).

7 Vale, P. F. & Little, T. J. CRISPR-mediated phage resistance and the ghost of coevolution past. *Proc Biol Sci* **277**, 2097-2103, doi:10.1098/rspb.2010.0055 (2010).

8 Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).

9 McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Molecular cell* **64**, 616-623, doi:10.1016/j.molcel.2016.08.038 (2016).

10 Nuñez, James K., Bai, L., Harrington, Lucas B., Hinder, Tracey L. & Doudna, Jennifer A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Molecular Cell* **62**, 824-833, doi:10.1016/j.molcel.2016.04.027 (2016).

11 Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science (New York, N.Y.)* **357**, 1113-1118, doi:10.1126/science.aao0679 (2017).

12 Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E5122-E5128, doi:10.1073/pnas.1618421114 (2017).

13 Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. *Nucleic acids research* **43**, 1749-1758, doi:10.1093/nar/gku1407 (2015).

14 Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nature Structural & Molecular Biology* **23**, 876-883, doi:10.1038/nsmb.3289 (2016).

15 Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137-141, doi:10.1038/nature24020 (2017).

16 Lee, D. H. & Schleif, R. F. In vivo DNA loops in araCBAD: size limits and helical repeat. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 476-480, doi:10.1073/pnas.86.2.476 (1989).

17 Wang, J. C. Helical repeat of DNA in solution. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 200-203, doi:10.1073/pnas.76.1.200 (1979).

18 Yoganand, K. N. R., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic acids research* **45**, 367-381, doi:10.1093/nar/gkw1151 (2017).

19 Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC genomics* **17**, 356-356, doi:10.1186/s12864-016-2627-0 (2016).

20 Eckweiler, D., Dudek, C.-A., Hartlich, J., Brötje, D. & Jahn, D. PRODORIC2: the bacterial gene regulation database in 2018. *Nucleic acids research* **46**, D320-D326, doi:10.1093/nar/gkx1091 (2018).

21 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

22 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).

23 Richter, C., Gristwood, T., Clulow, J. S. & Fineran, P. C. In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System. *PloS one* **7**, e49549-e49549, doi:10.1371/journal.pone.0049549 (2012).

24 Rollins, M. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E5113-E5121, doi:10.1073/pnas.1616395114 (2017).

25 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-W208, doi:10.1093/nar/gkp335 (2009).

26 Aeling, K. A. *et al.* Indirect Recognition in Sequence-specific DNA Binding by Escherichia coli Integration Host Factor. *Journal of Biological Chemistry* **281**, 39236-39248, doi:10.1074/jbc.m606363200 (2006).

27 Rice, P. A., Yang, S.-w., Mizuuchi, K. & Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. *Cell* **87**, 1295-1306, doi:10.1016/s0092-8674(00)81824-3 (1996).

28 Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. *Current Opinion in Structural Biology* **14**, 28-35, doi:10.1016/j.sbi.2003.12.003 (2004).

29 Mouw, K. W. & Rice, P. A. Shaping the Borrelia burgdorferi genome: crystal structure and binding properties of the DNA-bending protein Hbb. *Molecular Microbiology* **63**, 1319-1330, doi:10.1111/j.1365-2958.2007.05586.x (2007).

30 Kobryn, K., Naigamwalla, D. Z. & Chaconas, G. Site-specific DNA binding and bending by the Borrelia burgdorferi Hbb protein. *Molecular Microbiology* **37**, 145-155, doi:10.1046/j.1365-2958.2000.01981.x (2000).

31 Connolly, M. *et al.* Static Kinks or Flexible Hinges: Multiple Conformations of Bent DNA Bound to Integration Host Factor Revealed by Fluorescence Lifetime Measurements. *J Phys Chem B* **122**, 11519-11534, doi:10.1021/acs.jpcb.8b07405 (2018).

32 Petassi, M. T., Hsieh, S.-C. & Peters, J. E. Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons. *Cell* **183**, 1757-1771.e1718, doi:10.1016/j.cell.2020.11.005 (2020).

33 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225, doi:10.1038/s41586-019-1323-z (2019).

34 Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernández, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature* **577**, 271-274, doi:10.1038/s41586-019-1849-0 (2019).

35 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E7358-E7366, doi:10.1073/pnas.1709035114 (2017).

36 Vorontsova, D. *et al.* Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic acids research* **43**, 10848-10860, doi:10.1093/nar/gkv1261 (2015).

37 Wiegand, T. *et al.* Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. *The CRISPR journal* **3**, 378-387, doi:10.1089/crispr.2020.0069 (2020).

38 Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141-145, doi:10.1038/s41586-020-2018-1 (2020).

39 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318-356, doi:10.1016/s0022-2836(61)80072-7 (1961).

40 Pérez-Martín, J., Rojo, F. & de Lorenzo, V. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiological Reviews* **58**, 268-290, doi:10.1128/mr.58.2.268-290.1994 (1994).

41 Schleif, R. DNA LOOPING. *Annual Review of Biochemistry* **61**, 199-223, doi:10.1146/annurev.bi.61.070192.001215 (1992).

42 Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* **15**, 453-468, doi:10.1038/nrg3684 (2014).

43 Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science (New York, N.Y.)* **353**, aaf1175-aaf1175, doi:10.1126/science.aaf1175 (2016).

44 Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380-385, doi:10.1038/s41586-018-0569-1 (2018).

45 Nivala, J., Shipman, S. L. & Church, G. M. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nature microbiology* **3**, 310-318, doi:10.1038/s41564-017-0097-z (2018).

46 Kieper, S. N., Almendros, C. & Brouns, S. J. J. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiol Lett* **366**, fnz129, doi:10.1093/femsle/fnz129 (2019).

47 Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic acids research* **46**, 1007-1020, doi:10.1093/nar/gkx1232 (2018).

48 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research* **40**, 5569-5576, doi:10.1093/nar/gks216 (2012).

49 Kim, J. G., Garrett, S., Wei, Y., Graveley, B. R. & Terns, M. P. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic acids research* **47**, 8632-8648, doi:10.1093/nar/gkz677 (2019).

50 Kamashev, D. *et al.* Comparison of histone-like HU protein DNA-binding properties and HU/IHF protein sequence alignment. *PloS one* **12**, e0188037-e0188037, doi:10.1371/journal.pone.0188037 (2017).

51 Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic acids research* **48**, 748-760, doi:10.1093/nar/gkz1091 (2020).

52 Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic acids research* **42**, 8516-8526, doi:10.1093/nar/gku527 (2014).

53 Westra, Edze R. *et al.* Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology* **25**, 1043-1049, doi:10.1016/j.cub.2015.01.065 (2015).

54 Heussler, G. E., Miller, J. L., Price, C. E., Collins, A. J. & O'Toole, G. A. Requirements for Pseudomonas aeruginosa Type I-F CRISPR-Cas Adaptation Determined Using a Biofilm Enrichment Assay. *Journal of bacteriology* **198**, 3080-3090, doi:10.1128/JB.00458-16 (2016).

55 Kupczok, A., Landan, G. & Dagan, T. The Contribution of Genetic Recombination to CRISPR Array Evolution. *Genome biology and evolution* **7**, 1925-1939, doi:10.1093/gbe/evv113 (2015).

56 Gudbergsdottir, S. *et al.* Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Molecular microbiology* **79**, 35-49, doi:10.1111/j.1365-2958.2010.07452.x (2011).

57 Pul, Ü. *et al.* Identification and characterization ofE. coliCRISPR-caspromoters and their silencing by H-NS. *Molecular Microbiology* **75**, 1495-1512, doi:10.1111/j.1365-2958.2010.07073.x (2010).

58 Carte, J. *et al.* The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus. *Molecular microbiology* **93**, 98-112, doi:10.1111/mmi.12644 (2014).

59 Przybilski, R. *et al.* Csy4 is responsible for CRISPR RNA processing inPectobacterium atrosepticum. *RNA Biology* **8**, 517-528, doi:10.4161/rna.8.3.15190 (2011).

60 van Dijk, M. & Bonvin, A. M. J. J. 3D-DART: a DNA structure modelling server. *Nucleic acids research* **37**, W235-W239, doi:10.1093/nar/gkp287 (2009).

61 Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein science : a publication of the Protein Society* **27**, 14-25, doi:10.1002/pro.3235 (2018).

62 Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein science : a publication of the Protein Society* **30**, 70-82, doi:10.1002/pro.3943 (2021).

63 Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

64 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 research* **48**, D535-D544, doi:10.1093/nar/gkz915 (2020).

65 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

66 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

67 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

68 Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. *BMC bioinformatics* **8**, 312-312, doi:10.1186/1471-2105-8-312 (2007).

69 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490-e9490, doi:10.1371/journal.pone.0009490 (2010).

70 Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics* **69**, doi:10.1002/cpbi.96 (2020).

71 Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Molecular biology and evolution* **35**, 3041-3043, doi:10.1093/molbev/msy194 (2018).

72 Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28-36, doi:10.1111/2041-210x.12628 (2016).

73 Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *R package version 3.2.1* (2019).

74 Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686, doi:10.21105/joss.01686 (2019).

75 Qiu, D., Damron, F. H., Mima, T., Schweizer, H. P. & Yu, H. D. PBAD-based shuttle vectors for functional analysis of toxic and highly regulated genes in Pseudomonas and Burkholderia spp. and other bacteria. *Applied and environmental microbiology* **74**, 7422-7426, doi:10.1128/AEM.01369-08 (2008).

76 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)* **30**, 614-620, doi:10.1093/bioinformatics/btt593 (2014).

77 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421-421, doi:10.1186/1471-2105-10-421 (2009).

Supplemental Figures

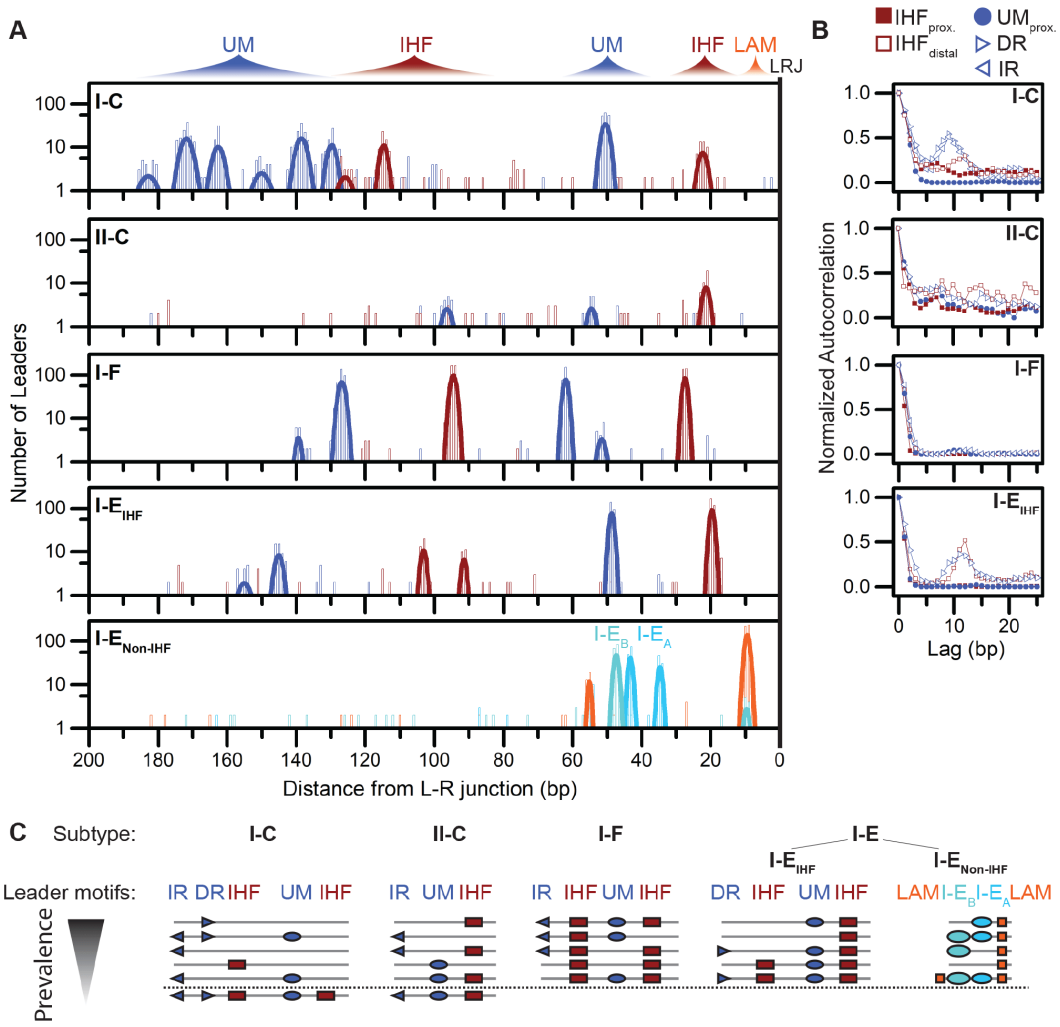


Figure S2.1. Naturally occurring indels preserve the phase of leader motifs. Related to Figures 2.2-2.4 and Table 2.1. (A) Tally of the midpoints of motifs found in I-C, II-C, I-F, I-E_{IHF}, I-$E_{Non\text{-}IHF}$ leaders. Phased motifs are fit to gaussian curves. (B) Autocorrelation plots of the midpoints of motifs found in I-C, II-C, I-F, I-E_{IHF}, I-$E_{Non\text{-}IHF}$ leaders. Common indels appear as peaks in the autocorrelation function. Motifs that are frequently shifted by 10-11 bp relative to the LRJ are retained in phase with the DNA helical repeat. (C) Schematics of observed leader motif architectures within each CRISPR subtype. IHF binding sites (IHF) are shown as red rectangles, upstream motifs (UMs) are shown in as blue ovals. The UM motif is often repeated, as a direct repeat (DR) and/or an inverted repeat (IR). The sequences of the UMs and the relative distances between leader motifs varies between leader subtypes. Many I-F leaders possess all four motifs, whereas most I-E leaders possess only the proximal UM and IHF site. I-C and II-C leaders with the full complement of leader motifs are rare, and not among the top five most prevalent leader architectures (below dotted line).

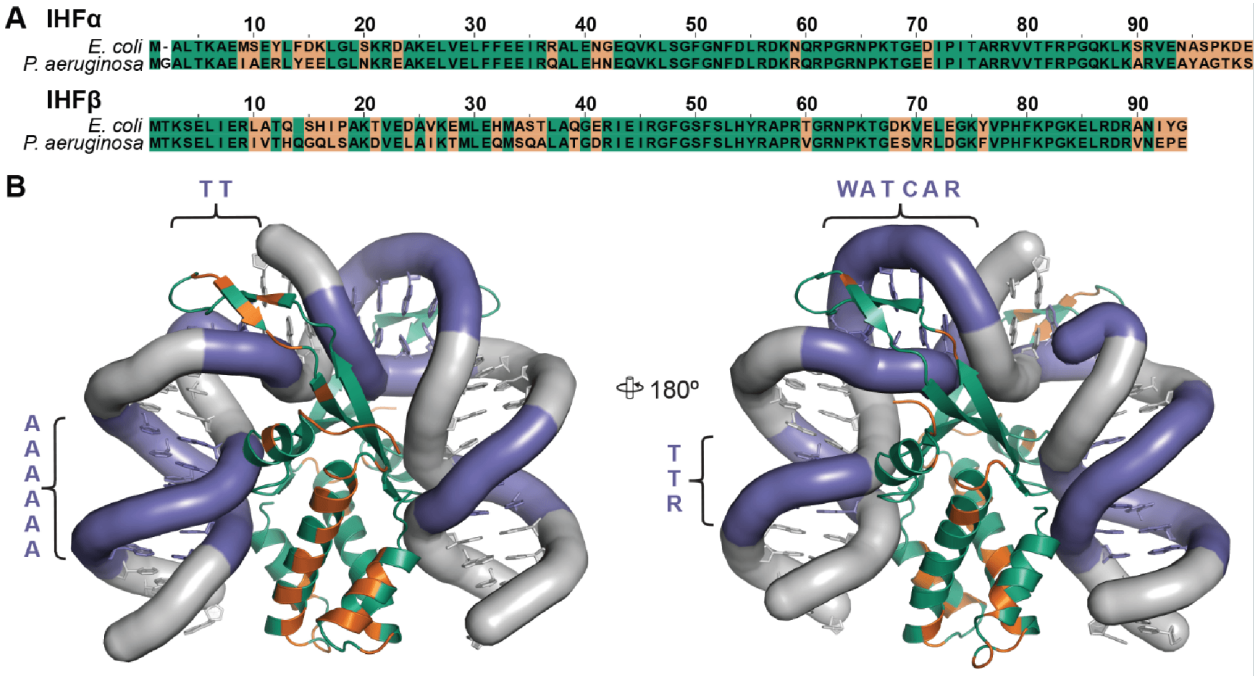


Figure S2.2. DNA-interacting residues are conserved between *E. coli* and *P. aeruginosa* IHF proteins. Related to Figure 2.3. (A) Pairwise alignments of *E. coli* and *P. aeruginosa* IHFα (top), and IHFβ (bottom). Identical residues (green) and non-identical residues (orange) are indicated. (B) Cartoon representations of *E. coli* IHF bound to cognate DNA (PDB:1IHF). DNA consensus sites (purple) are labelled. Non-consensus DNA regions are shown in light grey. Both IHF monomers are shown in green. Residues that are not identical in *P. aeruginosa* IHF monomers are shown (orange), none of which are expected to affect clamping of the A-tract, clamping of the TTR motif, kinking of DNA, or recognition of the TT and WATCAR motifs

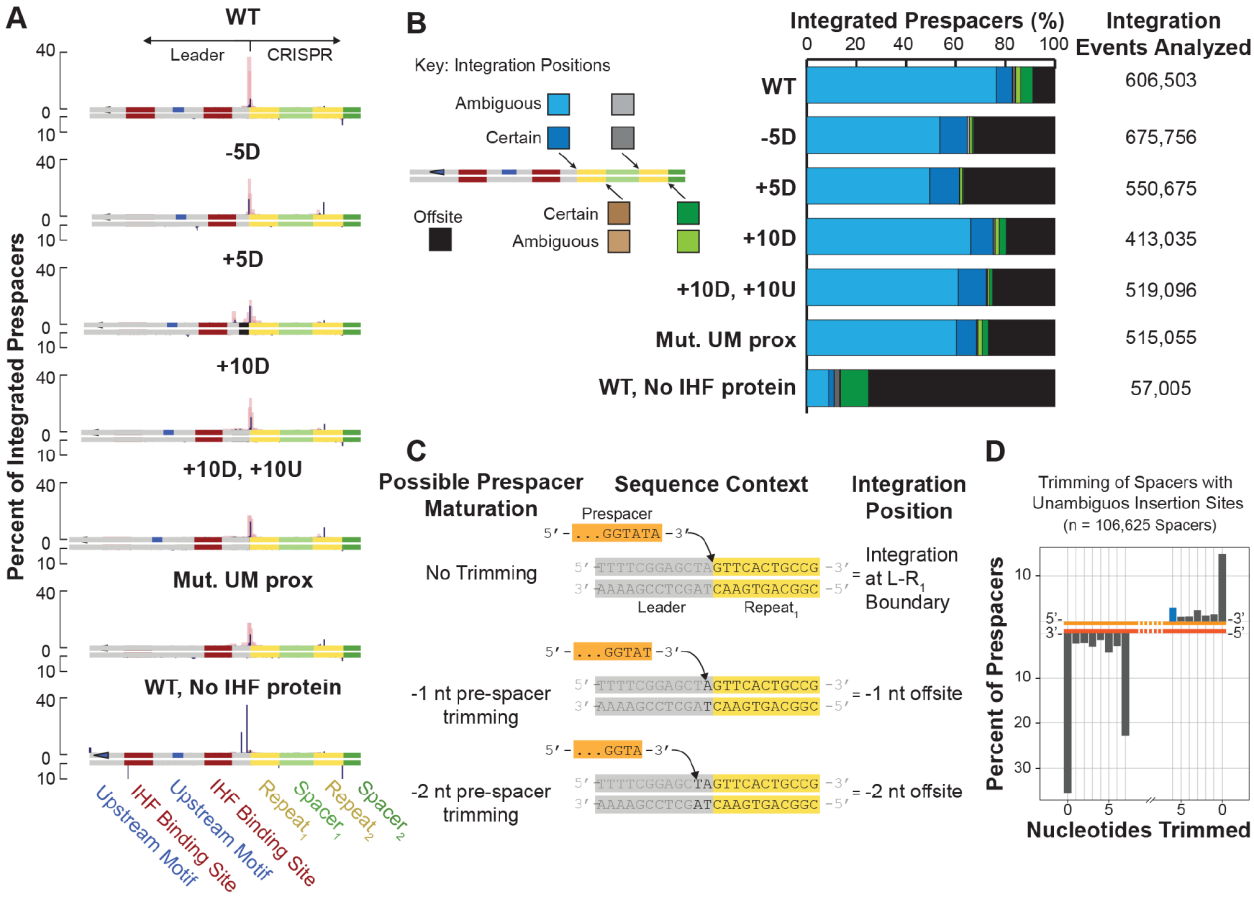


Figure S2.3. High-throughput sequencing and quantification of prespacer integration events. Related to Figures 2.3 and 2.4. (A) Graphical representation of Cas1-2/3-mediated integration events of a 40 bp dsDNA prespacer into a wildtype fragment of the CRISPR2 locus from P. aeruginosa PA14, in the absence of IHF (WT, No IHF protein), or in the presence of IHF (WT). Integration into variants containing leader mutations was also analyzed. "-5D" signifies deletion of 5 bp downstream of IHF proximal site. "+5D" signifies the addition of 5 bp downstream of IHF proximal site. "+10D" signifies the addition of 10 bp downstream of IHF proximal site "+10D,+10U" signifies the addition of 10 bp both downstream and upstream of the IHF proximal site."Mut.UM prox" is used to describe a CRISPR containing a scrambled proximal upstream motif. Prespacers with confirmed integration positions (blue) represent the minority of mapped reads, since most integration positions cover two to three nucleotides that could have come from either the prespacer or the CRISPR substrate (redboxes). Based on this uncertainty, we refer to these insertions as "ambiguous". (B) Data from panel A is quantified (right). Reads were counted as "ambiguous" insertion events at leader-repeat or spacer-repeat boundaries if the window of possible integration positions for that read covers one of the boundaries shown (left). (C)Ambiguous integration events arise because prespacers can be trimmed by Cas1-2/3 or contaminating nucleases prior to integration. An example of an ambiguous integration event at the leader-repeat1 boundary is shown. (D) Quantification of trimming events in prespacers confirmed integration positions into WT CRISPR substrate. Half of the prespacers analyzed are trimmed (49.8%), but there is no preference for cleavage at a PAM-spacer boundary (position

shown in blue) indicating that under the conditions tested, prespacer maturation may be stochastic.

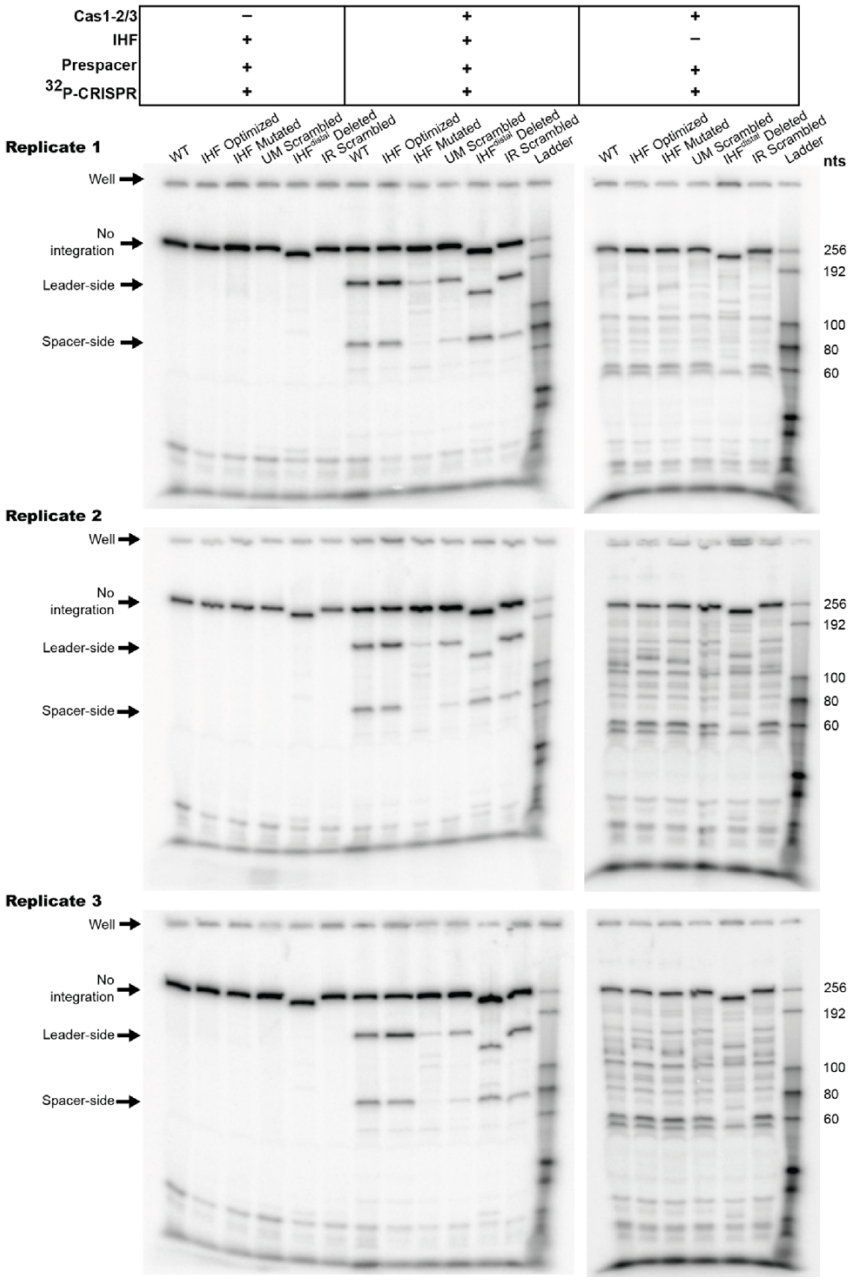


Figure S2.4. IHF binds I-F leaders to recruit an upstream motif needed for efficient prespacer integration. Related to Figure 2.3. Uncropped images of denaturing gel electrophoresis of integration reactions into a ^{32}P-labelled CRISPR fragment substrate. Reactions were performed in the absence of Cas1-2/3 (left), in the presence of all components (middle), or in the absence of IHF (right). Each row of gels depicts a replicate.

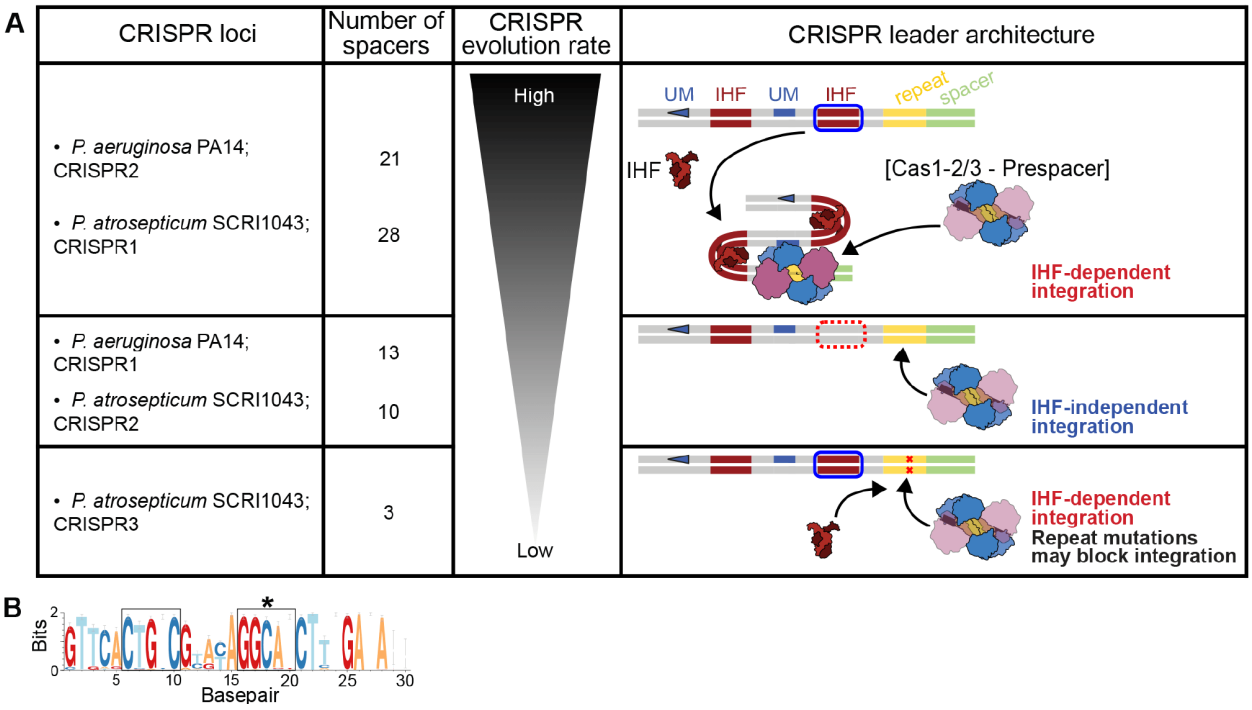


Figure S2.5. Summary of differential evolution rates between co-existing CRISPR loci. Related to Figure 2.3. (A) CRISPR loci found in *P. aeruginosa* PA14 and *P. atrosepticum* SCRI1043S6 fall into three categories based on the number of spacers, and rates of evolution. These categories correspond with three different CRISPR architectures. Rapidly evolving CRISPRs from these bacteria possess leaders with a proximal IHF site (midpoint 28 bp from the L-R junction), and a conserved upstream motif (midpoint 63 bp from the L-R junction) (top). IHF folds the leaders of these loci, to recruit an upstream motif to a specific surface of Cas1-2/3, facilitating efficient integration. Slower evolving CRISPRs do not have recognizable IHF binding sites in their leaders but do retain a correctly positioned upstream motif (middle). Integration at these loci is less efficient, and proceeds via an IHF-independent pathway. The slowest evolving CRISPR between these model organisms has the correct motifs in its leader, but integration may be blocked by a rare mutation within the inverted repeat of the first CRISPR repeat (bottom). Mutations in these regions of the CRISPR repeat are known to strongly inhibit integration. (B) Inverted repeat within I-F CRISPR repeat is highly conserved. Sequence logo of the first repeat in I-F CRISPRs. The highly conserved inverted repeats, necessary for efficient spacer integration, are boxed. The position of the C18T mutation that occurs in the first repeat of the *P. atrosepticum* CRISPR3 locus is indicated with an asterisk.

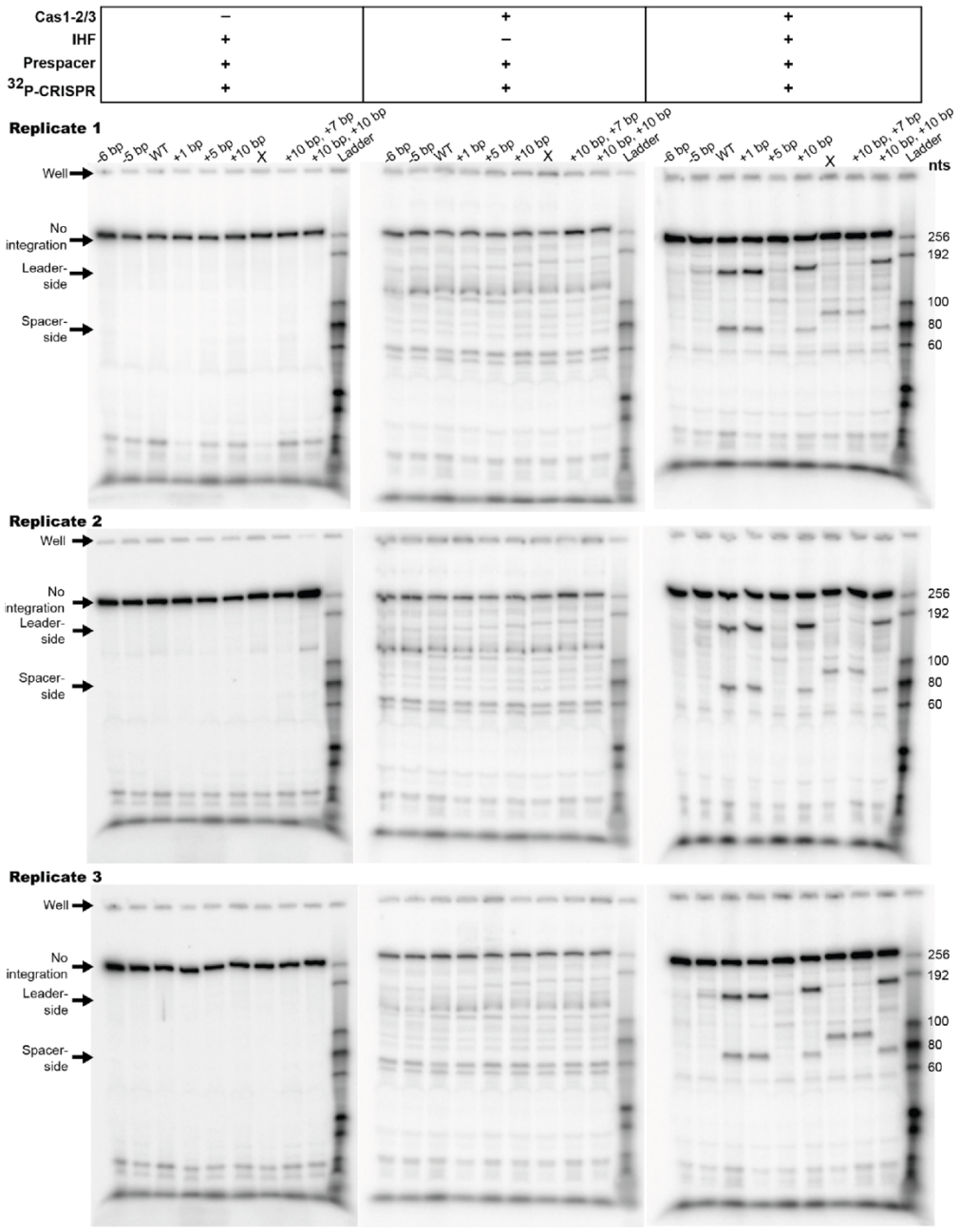


Figure S2.6. Leader motifs involved in CRISPR evolution are phase sensitive. Related to Figure 2.4. Uncropped images of denaturing gel electrophoresis of integration reactions into a ^{32}P-labelled CRISPR fragment substrate. Reactions were performed in the absence of Cas1-2/3 (left), in the absence of IHF (middle), or in the presence of all components (right). "X" marks lanes containing reactions that were not examined further and are not discussed in this paper. Each row of gels depicts a replicate.

Materials and Methods

Bacterial Strains

Escherichia coli DH5a (Thermo Fisher Scientific) cells were used to amplify plasmids used in this paper. *E. coli* BL21 DE3 (NEB) cells were used to express proteins used in this paper. *E. coli* were grown in LB (Lennox) media at either 37ºC, or 16ºC after induction of protein expression with 0.5 mM IPTG (isopropyl-b-D-thiogalactoside), shaking in baffled conical flasks at 200 rpm. The type I-F CRISPR system Cas proteins and CRISPR loci were cloned from *P. aeruginosa* UCBPP-PA14.

Modeling the I-F CRISPR integration intermediate complex

I-F CRISPR leaders contain an 8 bp insertion between the leader-repeat junction and the proximal IHF site relative to IHF site-containing I-E CRISPR leaders. To model the impact of this insertion, 8 bp of idealized B-form dsDNA containing 11 bp/turn was built using the 3D-DART web server.[60] This 8 bp duplex was inserted immediately downstream of the IHF proximal binding site in the I-E Cas1-2-IHF-CRISPR holo-complex structure (PDB: 5WFE) using PyMOL v1.8.2.3 (Schrödinger, LLC). Next, components of the I-F Cas1-2/3 adaptation complex were docked onto the I-E Cas1-2 adaptation complex. Docking was performed by superimposing the I-F Cas1 dimers (PDB: 3GOD) onto each I-E Cas1 dimer (PDB: 5WFE) using the “cealign” command in PyMOL v1.8.2.3 (Schro¨dinger, LLC). The Cas1 homodimers superimpose with a root mean squared deviation of 1.3 A˚over 144 α-carbon atoms. I-F Cas2/3 proteins (PDB: 5B7I) were similarly modeled by superimposing one copy of the I-F Cas2 domain of Cas2/3 onto each of the I-E Cas2 subunits. The Cas2 subunits superimpose with a root

mean squared deviation of 1.2 Å over 36 α-carbon atoms. Images of the resulting models were rendered using ChimeraX v1.1.[61,62]

Building a database of CRISPR leaders

A total of 15,567 complete bacterial genomes and 2,658 complete bacterial chromosomes were downloaded from the NCBI RefSeq Assembly database on June 10th of 2019. In addition, 351 complete archaeal genomes and 25 complete archaeal chromosomes were downloaded from the NCBI GenBank Assembly database on the same day. CRISPRDetect v2.419 was used to identify CRISPR loci in all downloaded genomes. Search parameters were set using the following command, "-word_length 11 -minimum_word_repeatation 3 -max_gap_between_crisprs 125 -repeat_length_cutoff 17 -minimum_repeat_length 23 -minimum_no_of_repeats 3 -check_direction 1 -array_quality_score_cutoff 3." The first repeat and 200 nucleotides upstream of each CRISPR locus (leader) were collected for downstream analyses. Regions corresponding to both sides of the CRISPR were downloaded when CRISPRDetect was not able to reliably determine the directionality (i.e., distinguish the leader from the trailer). Leaders were assigned to a particular CRISPR subtype by CRISPRDetect v2.4, or by proximity to subtype-specific *cas* genes. The remaining 13% of CRISPR loci (1,996 loci) could not be annotated by CRISPRDetect and lacked nearby *cas* genes.

Identification of conserved DNA motifs in CRISPR leaders

CRISPR leaders were analyzed using the MEME webserver with default settings.[25] Custom position weight matrices for I-E leader IHF binding sites, I-F leader IHF binding sites, I-E leader upstream motifs, I-F leader upstream motifs, I-C upstream motifs, II-C upstream motifs, $I\text{-}E_A$, $I\text{-}E_B$, and the I-E leader-anchoring motif are available in Data S1 (available at

doi.org/10.1016/j.cub.2021.05.068), and the FIMO-identified sequences found in CRISPR leaders have been submitted to the PRODORIC2 webserver and are available in Data S2 (available at doi.org/10.1016/j.cub.2021.05.068). A local copy of FIMO[21] was used to find significant matches to each of the position weight matrices. A minimum *p* value threshold of 1E-4 was used to define significant position weight matrices matches. The position weight matrix for the I-F leader IHF binding site does not represent the entire protein-occluded region of DNA. Therefore, tallies for the I-F IHF binding site were extended to represent the 29 bps of DNA that are occluded by bound IHF,[27] in order to calculate IHF binding site midpoints. This analysis allows for comparison of IHF binding site distributions between CRISPR leaders, and for comparison in the position of motifs mapped to the sense and anti-sense strands. Sequence logos for matched DNA sequences were generated with a local copy of WebLogo 3.7.[63]

The microbial genome dataset used here was downloaded from the NCBI RefSeq database on June 10th of 2019. As of April 1st of 2021, an additional 8,703 bacterial genomes, 1,277 bacterial chromosomes, 67 archaeal genomes and 6 archaeal chromosomes have been added to the NCBI RefSeq database. These additions represent about a third of the total number of currently available genomes and chromosomes, with large numbers of genomes originating from uncultured bacteria and the Phyla *Planctomycetes*, *Ignavibacteria* and *Candidatus Saccharibacteria*. The exclusion of these new genomes is not expected to change the conclusions of this paper, which focuses on CRISPR subtypes (I-E, I-F, I-C and II-C) predominantly found in *Actinobacteria*, *Proteobacteria* and *Firmicutes*.[64] However, we expect that additional bioinformatic studies will uncover more leader motifs that regulate CRISPR adaptation or transcription.

Phylogenetic analyses

CRISPR leaders containing sequences matching any of the position weight matrices described in this paper, were fetched using a Python v2.7 script that uses the Bio SeqIO package. A non-redundant list of CRISPR leaders was generated using CD-HIT v4.8.1[65,66] with a 95% identity cutoff. Non-redundant CRISPR leaders were then aligned with a local version of MAFFT v7.429,[67] using the following command-line options: "–genafpair–maxiterate 1000–thread 100–threadit 100–threadtb 100." The resulting alignment was then analyzed with MaxAlign v1.1[68] to find and remove misaligned or non-homologous sequences which introduced a large number of gaps in the alignment. The remaining leader sequences were then realigned as above. A maximum-likelihood phylogenetic tree was then generated from realigned leader sequences, using FastTree 2.1.11.[69] The tree was generated using the following parameters "-quote -gamma -spr 4 -mlacc 2 -slownni -nt." Trees were visualized in RStudio by overlaying the data of motif distances using ggtree[70-72] ggplot2[73] and tidyverse.[74]

Plasmid construction

The *cas1* and *cas2/3* genes from *P. aeruginosa* UCBPP-PA14 have been previously cloned into a spectinomycin-resistant p2S LIC vector[24] (Addgene, #89240). The *ihfA* and *ihfB* genes from *P. aeruginosa* UCBPP-PA14 were PCR-amplified to construct an N-terminal 6x-Histidine tagged variant of IHFα, an untagged IHFβ and an N-terminal Strep-II tagged variant of IHFβ. HRV3C protease cleavage sites were included between the affinity tags and the protein sequence of interest. 6x-His-IHFα and untagged IHFβ were cloned into MCS1 and MCS2 of pACYCDuet-1, respectively. This plasmid (pHisIHFαIHFβ) has been deposited (Addgene:

149384). StrepII-IHFβ was cloned into pET28a (pStrepIHFβ) and has been deposited (Addgene: 149385).

The *P. aeruginosa UCBPP-PA14* CRISPR2 locus (position 2935917-2937205, in the genome, GenBank: NC_008463), as well upstream leader DNA (311 bp) and downstream terminus DNA (303 bp) were PCR-amplified and cloned into pHERD30T,[75] (pCRISPR2_wt, Addgene: 149386). Leader sequence variants were generated via site directed mutagenesis using pCRISPR2 as template, followed by ligation of the linearized plasmids. All variants are deposited at Addgene, and are listed in the Key resources table (pCRISPR2_-6D, Addgene: 149387; pCRISPR2_-5D, Addgene: 149388; pCRISPR2_+1D, Addgene: 149389, pCRISPR2_+5D, Addgene: 149390; pCRISPR2_+10D, Addgene: 149391; pCRISPR_+10D+7U, Addgene: 149393; pCRISPR_+10D+10U, Addgene: 149394; pCRISPR2_IHF_Opt, Addgene: 149395; pCRISPR2_IHF_Mut, Addgene: 149396; pCRISPR2_motif_scram, Addgene: 149397; pCRISPR2_IHFdist_rm, Addgene: 162318; pCRISPR2_IRdist_scram, Addgene: 162319).

IHF expression and purification

E. coli BL21(DE3) cells were co-transformed with both pHisIHFα-IHFβ and pStrepIHFβ. Cells were grown in LB-Miller media (10 g/L Tryptone, 10 g/L NaCl, 5 g/L yeast extract), supplemented with 34 mg/mL Chloramphenicol and 50 mg/mL Kanamycin, at 37ºC and 200 rpm to an OD_{600} of 0.45. Cultures were then cooled on ice for one hour, without agitation, before the addition of 0.2 mM IPTG. Cells were then grown for an additional 18 hours at 16ºC, before centrifugation at 5000 *g* for 10 minutes. Cell pellets were suspended in Lysis Buffer (25 mM HEPES-NaOH pH 7.5, 500 mM NaCl, 10 mM Imidazole, 1mM TCEP, 5% Glycerol)

supplemented with 0.3x Halt™ Protease Inhibitor Cocktail (ThermoFisher), and sonication at 30% amplitude for a total on-time of 6 minutes (1 s on-time with 3 s off), at 4ºC. Lysate was clarified by two successive centrifugations at 12,000 rpm for 15 minutes each at 4ºC. Clarified lysate was then flowed over HisTrap HP resin (Cytiva), to affinity purify His-tagged IHF heterodimers. HisTrap HP resin was washed with 20 column volumes of IHF Lysis Buffer containing 50 mM Imidazole. His-tagged IHF heterodimers were then eluted from the HisTrap HP resin with Lysis Buffer containing 500 mM Imidazole. IHF was then concentrated at 4ºC (Corning Spin-X concentrators). 6x-Histidine and StrepII tags were cleaved from IHFα and IHFβ monomers respectively, in the presence of PreScision protease, overnight at 4ºC while dialyzing into Lysis Buffer containing no Imidazole. Remaining 6x-His-IHFα and PreScision protease were removed by affinity chromatography using HisTrap HP resin (Cytiva). The IHF heterodimer was concentrated and diluted in buffer to reduce the NaCl concentration to 200 mM. Untagged IHF heterodimer was then purified by affinity chromatography on Heparin Sepharose (GE Healthcare), which was washed with 10 column volumes of Heparin Wash Buffer (25 mM HEPES-NaOH pH 7.5, 200 mM NaCl, 5% Glycerol). IHF was eluted from HiTrap Heparin HP resin (Cytiva) with a linear gradient to Heparin Wash Buffer containing 2 M NaCl. Fractions containing IHF heterodimer were concentrated at 4º C (Corning Spin-X concentrators), before purification on a Superdex 75 size-exclusion column (Cytiva) in Heparin Wash Buffer.

Cas1-2/3 expression and purification

E. coli BL21(DE3) cells were transformed with pCas1Cas23 and grown in 2x 1 L LB-Miller media (10 g/L Tryptone, 10 g/L NaCl, 5 g/L yeast extract), supplemented with 50 mg/mL Spectinomycin, at 37ºC and 200 rpm to an OD_{600} of 0.45. Cultures were then cooled on ice for

two hours, without agitation, before being induced with 0.2 mM IPTG. Cells were then grown for an additional 18 hours at 16º C, before being centrifuged at 5000 *g* for 10 minutes. Each cell pellet, originating from 1 L of culture, was resuspended in 20 mLs of Cas1-2/3 Lysis Buffer (50 mM HEPES pH 7.5, 500 mM KCl, 10% Glycerol) supplemented with 0.3x Halt Protease Inhibitor Cocktail (ThermoFisher). Cells were lysed via sonication, and lysate was clarified as above. StrepII-tagged Cas1-Cas2/3 was affinity purified on StrepTrap HP resin (GE Healthcare) and eluted with Cas1-2/3 Lysis Buffer containing 3 mM desthiobiotin (Sigma-Aldrich). Eluate was concentrated at 4ºC (Corning Spin-X concentrators), before purification over a Superdex 200 size-exclusion column (Cytiva) equilibrated in 10 mM HEPES pH 7.5, 500 mM KCl, and 10% Glycerol.

Nucleic acid preparation

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

TCTACATGGTCTAGGAAAAGGACTTGGACAAGGAGGTATA-3') and antisense (5'-TATACCTCCTTGTCCAAGTCCT TTTCCTAGACCATGTAGA-3') strands (Europhins). To make ^{32}P-labeled CRISPR integration substrates, primers (Forward primer: 5'-CCAATTGCCCGAAGCTTC-3'; Reverse primer: 5'-TCCAGAAGTCACCACCCG-3') (Europhins) complementary to far upstream in the leader and to within the 2nd spacer of the CRISPR loci were used to amplify a DNA fragment containing most of the leader and the beginning of the CRISPR locus, from pCRISPR plasmid variants (deposited with Addgene) as templates. These PCR products were purified on a 2% (w/v) agarose native gel and extracted with a gel DNA recovery kit (Zymo Research). 1 pmole of dsDNA, corresponding to 2 pmoles of 50 ends, was end-labeled on both strands with 4 pmoles of [γ-^{32}P]ATP (PerkinElmer) by polynucleotide kinase (NEB) in 1x PNK buffer at 37ºC for 45 minutes. PNK was heat-denatured by incubation at 65ºC for 20 minutes. Spin column purification (G-25, GE Healthcare) was used to remove unincorporated radioactive nucleotides and to buffer exchange DNAs into 1x TE (10 mM Tris-HCl pH 8, 1 mM EDTA).

In vitro integration assays

Integration reactions were performed using 300 nM of splayed prespacer (purified as described above), 200 nM Cas1-2/3, ~1 nM of ^{32}P-labeled CRISPR variant fragment and 350 nM IHF, in Integration Buffer (20 mM HEPES pH 7.5, 150 mM Potassium acetate, 5 mM $MnCl_2$, 1 mM DTT, 5% Glycerol), for 1 hour at 37ºC. Reactions were stopped by the addition of SDS to 1% and subsequent phenolchloroform extraction. The nucleic acid containing layer was mixed 1:1 with 2x formamide loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) and denatured at 95ºC for 5 minutes, before resolving full-length ^{32}P-

labeled CRISPR strands from those fragmented by an integration event on a 8% (w/v) (29:1 mono:bis) polyacrylamide Urea gel in 1x TBE. Gels were dried and quantified using a Typhoon phosphorimager (GE Healthcare). The intensity of full-length CRISPR variant, leader-side integration fragments, spacer-side integration fragments, non-specific integration events, and background readings were quantified with Multi Gauge v3 (Fujifilm). We then calculated the no integration, leader-side integration, spacer-side integration and non-specific integration events, as percentages of total events.

High-throughput sequencing of in vitro integration products

Integration reactions were performed using 200 nM of a 40 bp dsDNA prespacer (purified as above), 200 nM Cas1-2/3, 1 nM of CRISPR variant fragment, and 350 nM IHF, in Reaction Buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 5 mM $MnCl_2$, 1 mM DTT, 5% Glycerol), for 1 hour at 37ºC. Reactions were then phenol-chloroform extracted, and then further purified with a DNA Clean and Concentrator Kit (Zymo Research). 1 mL of eluted DNA was used as a PCR template. Combinations of four primers (P1, 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCACCCGGCTTTCTTAG-3'; P2, 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAATTGCCCGAAGCTTC-3'; P3, 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGTCTAGGAAAAGGACTTGGAC-3'; P4, 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCCTTGTCCAAGTCCTTTTCC-3') were used to amplify all possible integration products and to simultaneously add Illumina adaptor sequences, using Q5 DNA polymerase (NEB). These PCR products were purified on a

1% (w/v) agarose gel and extracted with a gel DNA recovery kit (Zymo Research). DNA barcoding and paired-end sequencing were performed at the University of Montana Genomics Core, on an Illumina MiSeq 300 V2. Paired end reads were merged in PEAR[76] and aligned to integrated and CRISPR substrates in BLAST+.[77] Downstream sequence analysis was performed in RStudio. The region flanking an integration site often contained 1-3 nucleotides of homology to the end of a prespacer substrate, making it impossible to distinguish if the matching nucleotides came from the prespacer molecule or the CRISPR locus-containing substrate. In the latter event, the prespacer may have been trimmed prior to integration as was seen for about half of the spacers with a clearly discernible integration site. These integration events were therefore marked as 'ambiguous', with the range of true integration points spanning up to a three-nucleotide window (Figure S2.3).

Quantification and Statistical Analysis

A threshold *p* value of 1E-4 was used to report significant matches of motif position weight matrices to leader sequences by FIMO.[21] Autocorrelation analysis of leader motifs was performed in OriginPro (OriginLab). Quantification of gel bands for integration experiments was performed in Multi Gauge (FUJIFILM) image analysis software, from 3 independent reactions and denaturing PAGE gel images. The mean ± 1 standard deviation is reported. In Figure 2.3, errors were propagated to report the fold integration measured in CRISPR variants relative to wild-type. In Figure 2.4 the raw mean was plotted ± 1 standard deviation and the data was fit to a sine wave equation of $y = y_0 + A * \sin(\pi \frac{x - x_c}{w})$ (in which y_0 is the y axis offset, A is the amplitude, w is the period, and x_c is the phase shift). The R^2 of the fits were 0.95 for the leader-side integration data and 0.75 for the spacer-side integration data.

CHAPTER THREE

STRUCTURES AND STRATEGIES OF ANTI-CRISPR-MEDIATED IMMUNE SUPPRESSION

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Tanner Wiegand

Contributions: wrote the manuscript

Co-Author: Shweta Karambelkar

Contributions: wrote the manuscript

Co-Author: Joseph Bondy-Denomy

Contributions: wrote the manuscript

Co-Author: Blake Wiedenheft

Contributions: wrote the manuscript

Manuscript Information

Tanner Wiegand, Shweta Karambelkar, Joseph Bondy-Denomy, and Blake Wiedenheft

Annual Review of Microbiology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Annual Reviews
Volume 74
September 2020
Pages 21-37
DOI: 10.1146/annurev-micro-020518-120107

STRUCTURES AND STRATEGIES OF ANTI-CRISPR-MEDIATED IMMUNE SUPPRESSION

Tanner Wiegand,[1] Shweta Karambelkar,[2] Joseph Bondy-Denomy,[2,3] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Department of Microbiology and Immunology and Quantitative Biosciences Institute, University of California, San Francisco, California 94143, USA. [3]Innovative Genomics Institute, Berkeley, California 94720, USA

Abstract

More than 50 protein families have been identified that inhibit CRISPR (clustered regularly interspaced short palindromic repeats)-Cas-mediated adaptive immune systems. Here, we analyze the available anti-CRISPR (Acr) structures and describe common themes and unique mechanisms of stoichiometric and enzymatic suppressors of CRISPR-Cas. Stoichiometric inhibitors often function as molecular decoys of protein binding partners or nucleic acid targets, while enzymatic suppressors covalently modify Cas ribonucleoprotein complexes or degrade immune signaling molecules. We review mechanistic insights that have been revealed by structures of Acrs, discuss some of the trade-offs associated with each of these strategies, and highlight how Acrs are regulated and deployed in the race to overcome adaptive immunity.

Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and associated *cas* genes are essential components of diverse adaptive immune systems that defend bacteria and archaea from infection by foreign genetic elements. These immune systems are partitioned into two classes that have evolved independently but have been exchanged horizontally across taxa.[1,2] Class1 systems are divided into three types (I, III, IV) and 18 subtypes (A, B, C, etc.), but all class 1 systems consist of a multi-subunit RNA-guided surveillance complex.[2,3] Similarly, class 2 systems are divided into three types (II, V, VI) and 26 subtypes, but all class 2 systems consist of a single-protein effector that is guided by a CRISPR RNA (crRNA).[2,3] Despite the phylogenetic and functional diversity of these systems, they all seem to participate in defense.

Considerable effort has been dedicated to understanding how Cas proteins integrate fragments of foreign DNA at one end of the CRISPR locus, and how CRISPR DNA is transcribed and processed into short crRNAs that guide Cas nucleases to the DNA or RNA of invading parasites (Figure 3.1). Progress in this field has been frenetic and numerous reviews dedicated to mechanisms of CRISPR adaptation,[4-7] crRNA biogenesis[8,9] and interference[10-12] are available. As sophisticated and diverse as these immune systems are, phages and other genetic parasites have evolved mechanisms to neutralize these immune systems. Originally discovered in 2013, anti-CRISPRs (Acr) appear to mirror the diversity of the CRISPR systems themselves.[13-15] Much like CRISPR systems, anti-CRISPRs have attracted considerable attention and several reviews have recently been published that address Acr function, evolution, and methods of discovery.[16-19] Here we complement existing reviews by first introducing recent work on the expression and regulation of Acrs. The implications of *acr* regulation are discussed in the context

of CRISPR-Cas expression, before we move on to a discussion focused on the structures of Acr proteins and what they have taught us about the vulnerabilities of CRISPR RNA-guided defense systems.

Timing is everything

How do crRNA-guided surveillance complexes find complementary targets in a crowded intracellular environment, and on a time scale that affords protection from an invading virus that will (in some cases) program the cell for lysis (i.e., death) in the first few minutes after infection? The mechanism(s) that explain effective surveillance are complicated, but we know that detection of a complementary DNA target does not initially rely on unwinding the dsDNA duplex, which would be slow and energetically expensive.[20,21] Instead, detection of invading dsDNA (crRNA-guided detection of RNA relies on alternative mechanisms), starts with the identification of a short-duplexed sequence motif called a PAM (protospacer adjacent motif).[21-23] PAM binding is thought to destabilize the duplex and thereby facilitate crRNA-guided strand invasion.[21,24,25] If the adjacent sequence is not complementary, then the interaction is ephemeral, and the search continues.[26,27] In contrast, a PAM with an adjacent complementary sequence triggers a conformational change that activates the nuclease and prompts target destruction.[28-34] Collectively, this is an efficient process and some crRNA-guided surveillance complexes (e.g. Cas9 and Cascade) are predicted to find their targets in less than a minute.[27,35,36]

But if crRNA-guided immune surveillance is rapid and efficient, then how do phages escape detection and elimination? DNA mutations and modifications play an important role in phage escape,[37-40] but here we focus on the delivery of immune suppressors. To effectively suppress the immune system, phages must quickly produce or deliver Acrs before the genome is

identified and destroyed by crRNA-guided immune complexes.[41,42] It is conceivable (maybe even probable) that some phages package and inject Acr proteins along with their genomes. In fact, Stone *et al* recently determined the structure of a phage decoration protein that is a structural homolog of AcrIIC1.[43] Although this structural similarity raises the intriguing possibility that proteins associated with the virion could serve as readymade CRISPR-Cas antidotes, recent evidence now indicates that *acrs* are among the first genes to be transcribed and translated during an infection (Figure 3.1).[44,45]

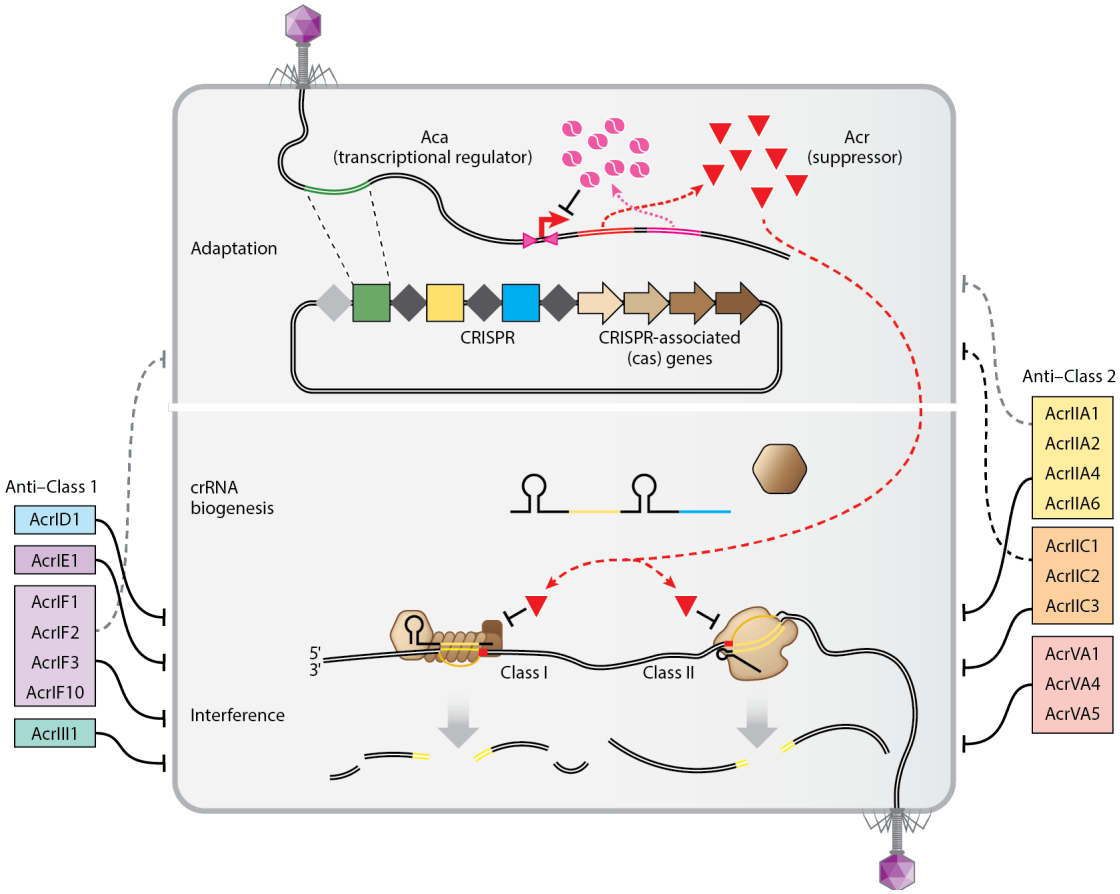


Figure 3.1. CRISPR-Cas adaptive immunity and structurally determined anti-CRISPRs (Acrs). Homodimeric Acr-associated (Aca) proteins (pink) repress Acr (red triangles) expression by binding to an upstream inverted repeat. Most suppressors of class 1 (left) and 2 (right) immune systems target the CRISPR RNA (crRNA)-guided surveillance complex and block DNA binding or nuclease activation.

To date, nearly fifty families of unique Acrs (some containing just a single known homolog) have been experimentally demonstrated to suppress type I, II, III or V CRISPR immunity (Figure 3.1 and 3.2).[46] These proteins are small (52-333 amino acids) and diverse, sharing little to no sequence similarity with other proteins.[14,15] The small size and diverse sequences make it difficult to identify new Acrs using standard homology-based search methods. However, Pawluk *et al.* first identified a conserved gene with a helix-turn-helix (HTH) motif that is encoded downstream of known anti-CRISPR genes, but is absent in related phages lacking anti-CRISPRs.[47] These anti-CRISPR associated (*aca*) genes have become effective new tools that serve as genetic landmarks for locating new Acrs, but until recently their functions have gone unreported. Stanley *et al.* recently demonstrate that *acr* and *aca* genes are immediately transcribed as a single RNA from the upstream promoter (i.e., polycistronic).[45] Collectively, this work, and a paper from Birkholz *et al.,* now show that Aca proteins are homodimers that repress expression of the operon by binding to inverted repeats in the promoter (Figure 3.1).[44,45] This results in a temporally controlled negative feedback loop that helps explain how phages deliver an early dose of Acrs, without the detrimental effects of runaway gene expression that occur in the absence of the repressor.[44,45] However, since Acr delivery appears to require transcription and translation (as opposed to delivery of the proteins directly) there is an intrinsic delay, and this delay provides an initial advantage to a previously "vaccinated" cell containing bespoke crRNA-guided complexes targeting that phage.[41,42] Borges *et al.* and Landsberger *et al.* recently demonstrated that infections of cells containing a CRISPR system targeting that phage are typically cleared, but each infection delivers small doses of Acr proteins that temporarily immunocompromise the cell. Thus, at high viral titers, Acrs accumulate to a critical intracellular

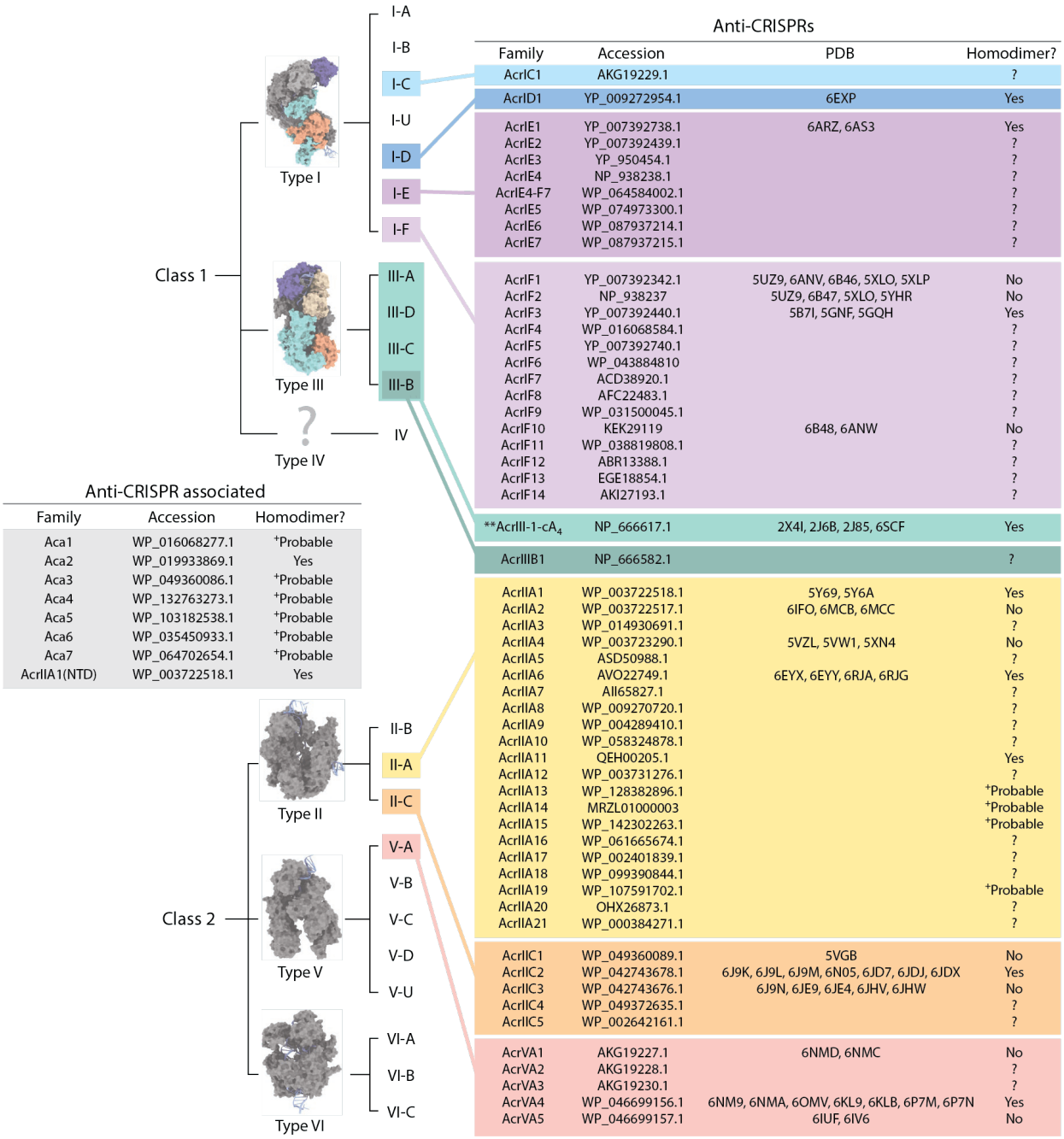


Figure 3.2. Anti-CRISPR (Acr) and anti-CRISPR-associated (Aca) proteins. Acr proteins that have been experimentally demonstrated to prevent CRISPR interference, with representative protein accession numbers, Protein Data Bank (PDB) identifiers, and propensity to form homodimers. Proteins with helix-turn-helix domains, but not confirmed to form homodimers, are listed as probable. Double asterisks (**) indicate Acrs with a suggested naming convention that includes the secondary messenger.

threshold that eventually overwhelms the immune system.[41,42] Remarkably, the intracellular threshold necessary for immunosuppression differs between Acrs and correlates with the Acr binding affinity (i.e., weak binders require higher Acr concentrations and vice versa).[41,42]

While the role of Acrs in blocking interference in previously vaccinated cells has been well established, very little work has been done to understand how Acrs impact adaptation or crRNA biogenesis (Figure 3.1). With the exception of a single paper, indicating that Acrs inhibit new sequence acquisition in type I-F systems,[48] no work has been published on the role of Acrs in processes upstream of interference. This is especially surprising since Cas9 plays a critical role in both interference and new sequence adaptation.[49,50] We anticipate that Acr proteins play an underappreciated role in limiting the efficiency of new sequence integration and that efforts to interrogate this aspect of the biology will reveal new insights into the CRISPR-anti-CRISPR dynamic.

Stoichiometric Inhibitors of CRISPR Defense

Anti-CRISPR Proteins that Masquerade as dsDNA

Parasites routinely use molecular mimicry to evade host immune responses.[51,52] Since many of the CRISPR-Cas systems target double-stranded DNA (dsDNA) (types I, II, and V), it is not surprising that phages have evolved a diverse repertoire of Acr proteins that serve as dsDNA decoys, which intercept the immune systems and prevent detection of invading DNA.[53-60] However, if dsDNA mimicry is a shared mechanism for diverse Acrs, then why does each Acr only interact with a specific surveillance complex? To address this question, we compared Acr structures to idealized B-form DNA and to the distorted DNA bound by each of the corresponding surveillance complexes (Figure 3.3A and D). These comparisons reveal a negative

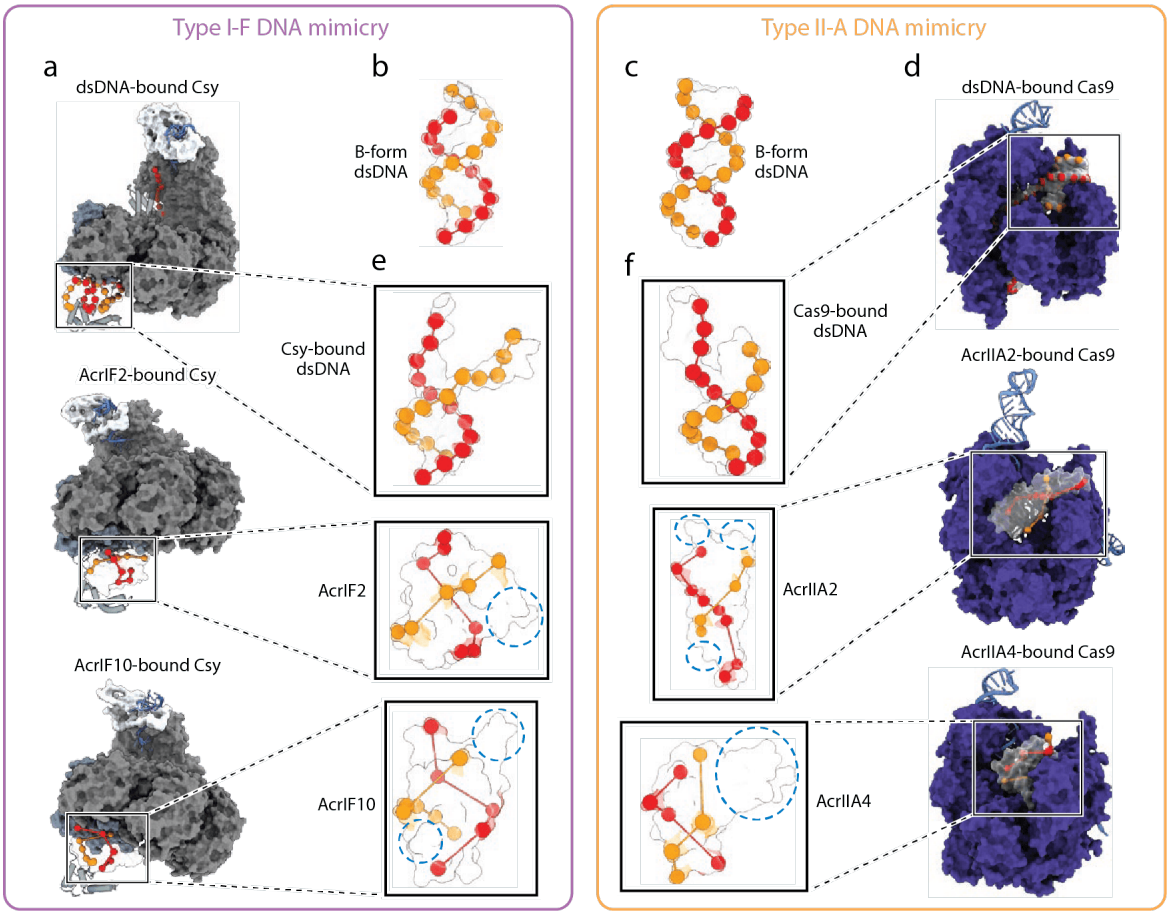


Figure 3.3. Anti-CRISPRs more closely mimic bound double-stranded DNA (dsDNA) than idealized B-form DNA. (A) Surface representation of Csy bound to dsDNA, AcrIF2, and AcrIF10 [Protein Data Bank (PDB): 6NE0, 5UZ9, 6B48]. (B, C) Surface representation of modeled B-form DNA with orbs highlighting phosphate groups of the DNA backbone. (D) Surface representation of Cas9 bound to dsDNA, AcrIIA2, and AcrIIA4 (PDB: 4UN3, 6MCB, 5VW1). (E, F) Surface representation of DNA bound by surveillance complex with orbs highlighting phosphate groups of the DNA backbone (Top) and surface representations of bound anti-CRISPRs with orbs highlighting pseudohelical arrangement of acidic residues and auxiliary structural features (dashed blue circles in bottom two insets).

charge distribution on Acrs that more closely resembles the pattern of phosphates on DNA bound by the surveillance complex than idealized B-form (Figure 3.3B-C and E-F). In addition to the pseudo-helical presentation of negatively charged residues, we found that Acrs disguised as (bound) dsDNA are also often adorned with additional structural features (Figure 3.3E-F). These

ancillary structural "decorations" often obscure the helical charge distribution characteristic of dsDNA, but they are expected to play important roles in affinity and specificity of Acr target selection. Collectively, the shape and charge distribution of a specific Acr, as well as the additional structural adornments that are unique to each Acr, may help explain why each DNA mimic exclusively targets the surveillance complex of a specific subtype.

Acr-Mediated Dimerization

Anti-CRISPRs that target diverse immune subtypes often form homodimers.[61-71] This strategy may increase the effective size of an otherwise small Acr, which may in turn increase the affinity and specificity of target interactions. Moreover, homodimeric-Acrs often dimerize their corresponding immune system targets (e.g., AcrIIA6 and AcrVA4), which may balance the costs of making more Acrs upon infection (i.e., two Acrs neutralize two Cas). AcrIIA6 forms a stable homodimer that recognizes one molecule of Cas9 through a series of high-affinity interactions contributed by each Acr subunit (Figure 3.4A).[63-65] Residues that produce interactions on one face of the homodimer are symmetrically presented on the opposing face and these residues are free to engage with an additional molecule of Cas9. Thus, one homodimer neutralizes two molecules of Cas9, and formation of the homodimeric Acr simultaneously increases the affinity for both molecules of the Cas9 target.

The benefits of homodimer formation are less clear in the case of AcrVA4. In contrast to the coordinated binding interactions by AcrIIA6, one subunit of an AcrVA4 dimer forms the majority of contacts with a single molecule of Cas12 (Figure 3.4B).[66-68] Theoretically, a

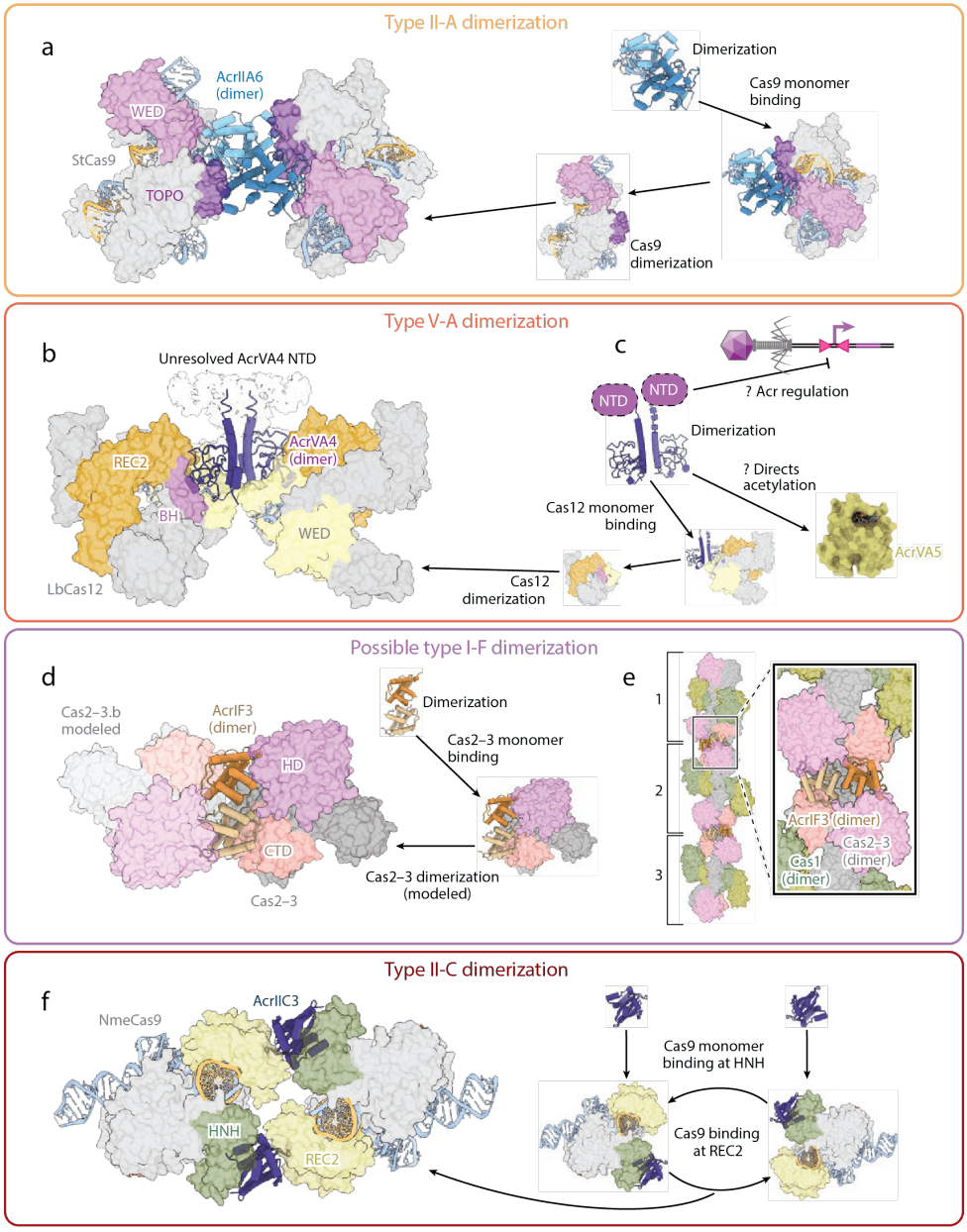


Figure 3.4. Acrs that oligomerize Cas protein targets. (A, B) AcrIIA6 and AcrVA4 dimerize StCas9 and LbCas12, respectively (PDB: 6RJA, 6NM9; EMDB: 9398). (C) AcrVA4 homodimer formation may facilitate *acr* regulation and/or guide AcrVA5 to the appropriate target (PDB: 6NMA, 6IUF). (D) Model for AcrIF3-mediated oligomerization of the Cas1–2/3 integration complex of the type I-F immune system. The AcrIF3 dimer interacts with the HD nuclease domain and CTD of Cas2/3 (PDB: 5B7I). (E) Proposed model for AcrIF3-mediated dimerization would enable bidirectional oligomerization of Cas1–2/3 heterohexamers, three of which are illustrated. Modeled Cas1–2/3 supercomplex joined by AcrIF3 (PDB: 5B7I, 3GOD; EMDB: 8558). (F) Two AcrIIC3 molecules join two molecules of NmeCas9. Domains of contact colored (PDB: 6JE4). Abbreviations: ACR, anti-CRISPR; BH, bridge-helix domain; CTD, carboxy

terminal domain; EMDB, Electron Microscopy Data Bank; HD, histidine-aspartate nuclease domain; HNH, HNH nuclease domain; NTD, N terminal domain; PDB, Protein Data Bank; REC2, recognition domain; TOPO, topoisomerase-homology domain; WED, wedge domain.

monomer of AcrVA4 would be just as effective as the dimer. Consistent with this idea, the flexible N-terminal domain (NTD) of AcrVA4 (unresolved in Cryo-EM structures) was shown to be required for Acr dimerization but dispensable for Cas12 inhibition.[66] However, additional selective pressures may be at play. Dimerization may be critical for regulating expression of *acrVA4*. Since the anti-CRISPR is not flanked by a known *aca* gene but does occur immediately downstream of two inverted repeats (IRs),[72] AcrVA4 may be regulated through an *aca*-like mechanism. While the role of the AcrVA4 NTD remains unknown, work recently deposited on Bioarchive (BioRxiv) and not yet peer reviewed, shows that the NTD regions of some Acrs (e.g., AcrIIA1) are not required for immune system inhibition, but rather function as Aca-type regulators of *acr* expression.[73] In fact, the NTD of AcrIIA1 is also required for dimerization,[70] again suggesting that while an Acr may form a dimer, this is not always necessary for Acr function. The NTD of AcrVA4 may play a similar role in regulating expression of the Acr, and this function may be dependent on dimerization of the suppressor. An alternate possibility for the biological role of AcrVA4 dimerization is that two molecules of the anti-CRISPR may be needed for interactions with AcrVA5. These Acrs cooccur in phage genomes and have been reported to form a complex,[72] however this assembly dissociates upon AcrVA4-LbCas12 binding.[68] Since AcrVA5 is an enzymatic inhibitor with little substrate specificity (see below), then its association with AcrVA4 may impart target specificity on an otherwise non-specific enzyme.

Unlike Acrs that dimerize the crRNA-guided surveillance complex, AcrIIC2 forms a homodimer but fails to dimerize Cas9, suggesting two molecules of the suppressor are needed

for each Cas target.[63,64] The AcrIIC2 dimer binds to apo-Cas9 (i.e., sans RNA) and prevents crRNA (or sgRNA) guide-loading, which may increase proteolytic degradation of newly expressed Cas9 proteins.[63,64] This makes AcrIIC2 one of the more confounding and potentially interesting Acrs, since this mechanism implies that it is unable to prevent interference by Cas9 that is already loaded with a guide. This Acr would have limited utility in a cell with preformed Cas9s carrying crRNAs targeting that phage. However, since Cas9 must associate with the crRNA to fulfill its role in new sequence acquisition,[49] then it is possible that AcrIIC2 serves as a suppressor that inhibits both interference and Cas9-dependent integration of foreign DNA into the CRISPR.[64]

Based on the emerging theme of homodimeric Acrs that trigger dimerization of their targets, we wondered if other Acr dimers might be capable of achieving a similar function. AcrIF3, forms a homodimer that binds the type I-F transacting nuclease/helicase (Cas2/3).[30,61,62,74] Structures of the AcrIF3 homodimer bound to Cas2/3 reveal a series of contacts made by each AcrIF3 molecule that are arranged along one face of the homodimer.[61,62] To determine if AcrIF3 might be capable of dimerizing Cas2/3, we modeled an additional molecule of Cas2/3 onto the crystal structure of dimeric AcrIF3 bound to Cas2/3 (Figure 3.4C). In this model, there are no major structural impediments that preclude AcrIF3-mediated dimerization of Cas2/3, and the majority of residues on the solvent exposed binding face of AcrIF3 are available to interact with a second molecule of Cas2/3. This suggests that AcrIF3 may tether two molecules of the helicase-nuclease to prevent interference. AcrIF3 has also been shown to block adaptation.[48] The mechanism for blocking adaption has not been well-established but Rollins *et al.* previously shown that AcrIF3 binds to Cas2/3 alone and to Cas2/3 in complex

with Cas1.[74].In most type I systems, Cas2 and Cas3 are separate proteins involved in adaptation and interference, respectively. However, in I-F systems, these proteins are fused into a single polypeptide that forms a large (~375 kDa) heterohexameric complex with Cas1 ($Cas1_4$-$Cas2/3_2$) (Figure 3.4D).[74-76] This suggests that AcrIF3 might be a more efficient immune suppressor than has previously been appreciated and its ability to block adaptation and interference, might be complemented by the ability to trigger oligomerization of the Cas1-2/3 complex.

Why Two Acrs?

Unlike the homodimeric Acrs that benefit from larger binding surfaces, and in many cases dimerize their targets, a few Acrs function as monomers that bind a single surveillance complex at multiple sites (e.g., AcrIF1 and AcrIIC3). The advantage of this 'strength in numbers' strategy is unclear, since multiple Acrs binding a single Cas protein will increase the anti-CRISPR concentration needed to saturate targets and slow down the process of immune suppression. For example, AcrIF1 blocks access to the crRNA-guide by binding to two different sites on the Csy complex,[53-55] but it is unknown if binding at both sites is necessary to prevent crRNA-guided target engagement. After PAM-binding, the crRNA is expected to directionally unwind the DNA duplex, starting at the PAM proximal end of the guide. Mismatches nearest the PAM (referred to as the seed region of the crRNA) prevent target binding, indicating that hybridization within the seed region of crRNA is critical to target recognition.[77,78] One of the two AcrIF1 molecules that bind the Csy complex block access to the crRNA seed. However, it is unclear if this binding site is sufficient to impede target binding or if a second molecule of AcrIF1 must also bind further up the crRNA backbone to prevent target hybridization.

The advantages of multiple binding sites is more obvious for AcrIIC3, which binds to both the recognition lobe (REC2) and the HNH-nuclease domain of Cas9.[34,63,79] In fact, AcrIIC3-mediated dimerization of Cas9 is necessary for suppression, as a single molecule of AcrIIC3 is unable to prevent Cas9 target cleavage.[63] Recently published structures reveal that two molecules of AcrIIC3 bridge oppositely oriented molecules of Cas9, creating a ring-shaped structure (Figure 3.4E).[34] AcrIIC3 prevents activation of the Cas9 HNH-nuclease,[34] but it is unclear why one Acr protein is unable to block this conformational rearrangement. To understand the mechanism of AcrIIC3, we superimposed a molecule of the Acr onto structures of Cas9 in multiple conformational states. This analysis reveals that Cas9 conformational changes induced by target binding and nuclease activation disrupt contacts with AcrIIC3, and these conformational changes must dislodge the anti-CRISPR in the absence of additional stabilizing contacts with a second molecule of Cas9. This indicates that assembly of the ring-shaped heterotetramer provides additional contacts to both molecules of AcrIIC3, allowing each suppressor to pin the HNH domains of the two Cas9s in inactive conformations. Collectively, this suggests a model where two AcrIIC3-Cas9 heterodimers bind one another to neutralize each surveillance complex (Figure 3.4E), and that a monomer of AcrIIC3 is insufficient to inhibit one molecule of Cas9 but two Acrs collectively inhibit two surveillance complexes.

Enzymatic Suppressors of CRISPR Immunity

Covalent Modification of the Surveillance Complex

In contrast to Acrs with steric and allosteric mechanisms of immunosuppression, AcrVA1 and AcrVA5 are enzymes that covalently modify the Cas12-crRNA complex, making them potent suppressors even at sub-stoichiometric concentrations.[68,69,80] This may be especially

beneficial for type V suppressors, since Cas12 is capable of multi-turnover targeting.[81] AcrVA1 is an endoribonuclease that cleaves between the fifth and sixth position of the crRNA (Figure 3.5A and B), while AcrVA5 is a GCN5-related N-acetyltransferase (GNAT) that acetylates lysines (Figure 3.5C and D).[68,69,80] Initially, AcrVA5 was thought to target acetylation to a specific PAM sensing residue (i.e., K635), but Dong *et al.* detected widespread acetylation of Cas12, which may indicate that AcrVA5 is more promiscuous than previously appreciated (Figure 3.5C). The biochemical promiscuity of AcrVA5 is consistent with the lack of an N-terminal specificity domain that is found in other GNAT homologs (Figure 3.5D).[82] However, untargeted acetylation seems dangerous and ineffective. The previously reported formation of a complex between AcrVA5 and AcrVA4,[68] suggests that AcrVA4 may direct the acetyltransferase activity of AcrVA5 to the appropriate (i.e., Cas12) substrate.

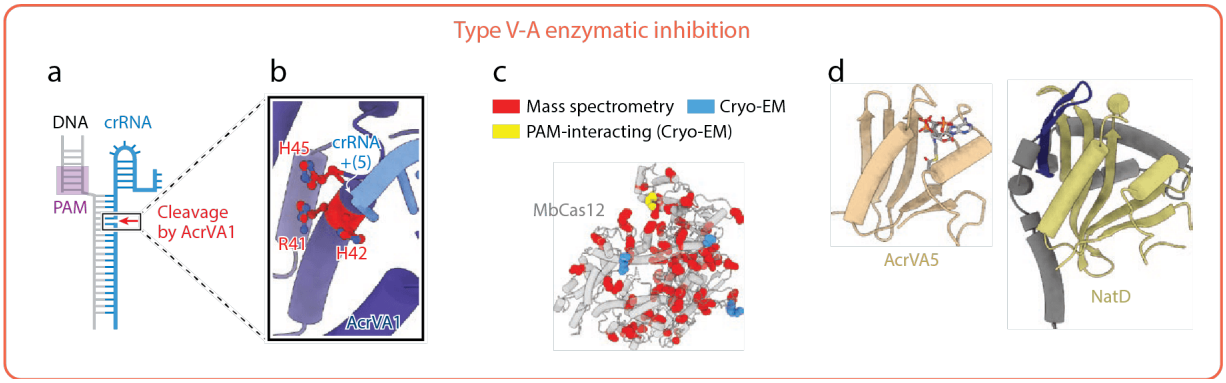


Figure 3.5. Enzymatic Acrs that inhibit type V-A CRISPR-Cas systems. (A) AcrVA1 is an endoribonuclease that cleaves between the fifth and sixth positions of the crRNA guide. (B) Active-site residues of AcrVA1 (red) responsible for crRNA cleavage (PDB: 6NMD). (C) Residues of MbCas12 acetylated by AcrVA5 according to mass spectrometry (red) or cryo-EM and mass spectrometry (blue and yellow) (PDB: 6IV6). (D) Comparison of AcrVA5 (tan) and closest structural homolog (NatD from *Homo sapiens*) (RMSD = 0.82 A˚ for 36 equivalently positioned C-α atoms). Structural features of NatD shared with AcrVA5 colored in olive. Two NatD β strands that determine acceptor substrate specificity (dark blue) and additional N-terminal structural features (gray) (PDB: 6IUF, 4U9W). Abbreviations: Acr, anti-CRISPR; crRNA, CRISPR RNA; EM, electron microscopy; PAM, protospacer-adjacent motif; PDB, Protein Data Bank; RMSD, root-mean-square deviation.

Degradation of Second Messenger Molecules

Cyclic oligonucleotides are important regulators of immune systems in all domains of life (25a, 45b).[83,84] Recently it has been demonstrated that detection of invading nucleic acids by the type III CRISPR-Cas immune systems triggers synthesis of cyclic-oligoadenylate (cA_4), which is a signaling molecule that activates additional nucleases that are essential for immune clearance.[85] The importance of cA_4 in immune signaling is highlighted by the recent discovery of a widespread anti-CRISPR (i.e., AcrIII-1) that enzymatically degrades cA_4, thereby neutralizing the immune response.[86] AcrIII-1, which has a novel fold, assembles into a homodimer that specifically recognizes the symmetry of cA_4. Though it may be a coincidence, we find the importance of symmetry imposed by homodimers to be remarkable. This includes homodimeric Acrs that dimerize the Cas proteins they target, homodimeric Acas that recognize the symmetry of inverted DNA repeats in the promoter, and now homodimeric Acrs that recognize the symmetry of second messengers.

In our opinion, AcrIII-1 represents an emerging ‘class’ of anti-CRISPRs that transcend traditional subtype boundaries because of their unique mechanisms of action. Given the lack of reliance on a specific protein-protein interaction, this anti-CRISPR is presumed to inhibit any system that relies on cA_4, regardless of subtype.[86] This suggests that, like Cas proteins, second messengers are also in conflict, and we expect that Acrs will mirror the diversity of the messengers themselves. At a practical level, Acrs that target second messengers, rather than a subtype specific protein, may defy the existing Acr naming scheme, which typically includes a letter to designate a specific subtype rather than a second messenger that might be present in many different subtypes. For this new generation of Acrs, we suggest a naming scheme that incorporates the specific messenger (e.g., AcrIII-1-cA_4).

The Future of Anti-CRISPR

Gazing into the anti-CRISPR crystal ball, it seems that only one thing is certain. The discovery of new anti-CRISPRs will continue to provide new insights about Acr mechanisms, while simultaneously revealing mechanistic vulnerabilities of the immune systems that they inhibit. To date, we are aware of 14 stoichiometric inhibitors that either block DNA binding (PAM recognition or hybridization to the guide) or allow binding but prevent nuclease activation. DNA mimics appear to be a common theme in the type I, II, and V systems, and we anticipate the discovery of Acrs that adopt similar strategies for RNA-targeting systems (type III and type VI). Protein mimics, like AcrIF3 (which mimics the helical bundle of Cas8f), also appear to be effective suppressors, and we expect that protein mimics will be identified with increasing frequency. Similarly, we anticipate that enzymatic inhibitors are currently underrepresented, and we expect that multiple-turnover enzymes that post-translationally modify Cas proteins or post-transcriptionally modify crRNAs will continue to reveal the versatility of these inhibitors.

As the repertoire of crRNA-guided nuclease continues to expand, the unique features of each system (i.e., PAM requirements, protein sizes, and cleavage mechanisms), are seamlessly incorporated into the growing quiver of genome editing tools that each have specialized chemical properties that make them more appropriate for particular applications.[85,87] Much like the CRISPRs, Anti-CRISPRs are now also being repurposed for applications in genome engineering. The most frequent use of Acr proteins has been as rheostats for tunable and tissue-specific control of gene editing or gene repression/activation.[88] Anti-CRISPRs have now been deployed using optogenetic and ligand-based post-translational Acr switches for control of Cas9 or dCas9

activity.[88,89] Additional proof-of-principle work has demonstrated the ability of anti-CRISPRs to reverse or limit Cas9-based gene drives and reduce off-target gene editing by limiting excessive Cas protein expression/activity through delayed Acr delivery,[59,90] or direct fusion of Acr mutants to the Cas9 nuclease.[91] We predict similar applications will emerge for Acrs that target Cas12- and Cas13-based systems, as well as Class I systems that have recently been domesticated for gene-editing,[83,92] or RNA-guided delivery of large DNA cargos to specific genetic addresses.[87,93]

Given the prevalence of CRISPR systems in archaeal genomes (~85%),[3,94] we expect that archaeal viruses will deploy the most diverse, abundant, and innovative Acr-based solutions for subverting CRISPR-based immunity and that these Acrs will in turn drive immunological innovation. Aside from AcrIII-1, only two other archaeal anti-CRISPR proteins have been described to date (i.e., AcrID1 and AcrIIIB1).[95,96] Remarkably, SIRV2 contains up to 12 copies of AcrID1.[96] This observation suggests that duplicated Acrs may be involved in dosage compensations, in a way that is conceptually similar to the expansions of K3L suppressors in poxviruses (DNA accordions).[97] Alternatively, or perhaps in addition, it is possible that each of these paralogs have nuanced, nonredundant functions.

The work on AcrIII-1 highlights the biological innovation that occurs in response to the selective pressures imposed by viruses and encourages us to think beyond the stoichiometric or enzymatic proteins that bind directly to Cas proteins or crRNAs. We anticipate that, like eukaryotic viruses, phages and viruses that infect archaea will also use noncoding RNAs or peptide-based inhibitors that serve as decoys or molecular jamming devices that block or redirect the immune systems. We are inspired by the recent work on Acr proteins that degrade cyclic nucleotides and we are reminded of recent work demonstrating that poxviruses also degrade

cyclic nucleotides that are essential for immune signaling in eukaryotic immune systems.[98] We anticipate that eukaryotic and prokaryotic immunologists will continue to identify common themes in immune defense and viral counter defense and based on historical precedence, we anticipate that many of these new discoveries will lead to disruptive new technologies that will, once again, transform industry and medicine.

Acknowledgements

B.W. is the founder of SurGene, LLC, and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. J.B.-D. is a scientific advisory board member of SNIPR Biome and Excision Biotherapeutics and a scientific advisory board member and co-founder of Acrigen Biosciences.

References

1 Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet*, 1-13 (2019).

2 Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biology direct* **6**, 38 (2011).

3 Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, doi:10.1038/s41579-019-0299-x (2019).

4 Amitai, G. & Sorek, R. CRISPR–Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* **14**, 67 (2016).

5 Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR-Cas systems. *Mol Cell* **61**, 797-808 (2016).

6 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:10.1126/science.aal5056 (2017).

7 McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat Rev Microbiol* **17**, 7-12 (2019).

8 Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS microbiology reviews* **39**, 428-441 (2015).

9 Charpentier, E., van der Oost, J. & White, M. F. in *CRISPR-Cas Systems* 115-144 (Springer, 2013).

10 Klompe, S. E. & Sternberg, S. H. Harnessing "A Billion Years of Experimentation": the ongoing exploration and exploitation of CRISPR–Cas immune systems. *The CRISPR journal* **1**, 141-158 (2018).

11 Hille, F. *et al.* The biology of CRISPR-Cas: backward and forward. *Cell* **172**, 1239-1259 (2018).

12 Plagens, A., Richter, H., Charpentier, E. & Randau, L. DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol Rev* **39**, 442-463 (2015).

13 Bondy-Denomy, J. *et al.* Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136-139, doi:10.1038/nature15254 (2015).

14 Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429-432, doi:10.1038/nature11723 (2013).

15 Pawluk, A., Bondy-Denomy, J., Cheung, V. H., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. *MBio* **5**, e00896, doi:10.1128/mBio.00896-14 (2014).

16 Borges, A. L., Davidson, A. R. & Bondy-Denomy, J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol* **4**, 37-59, doi:10.1146/annurev-virology-101416-041616 (2017).

17 Pawluk, A., Davidson, A. R. & Maxwell, K. L. Anti-CRISPR: discovery, mechanism and function. *Nat Rev Microbiol* **16**, 12 (2018).

18 Hwang, S. & Maxwell, K. L. Meet the Anti-CRISPRs: Widespread Protein Inhibitors of CRISPR-Cas Systems. *CRISPR J* **2**, 23-30, doi:10.1089/crispr.2018.0052 (2019).

19 Stanley, S. Y. & Maxwell, K. L. Phage-Encoded Anti-CRISPR Defenses. *Annu Rev Genet* **52**, 445-464, doi:10.1146/annurev-genet-120417-031321 (2018).

20 Santiago-Frangos, A., Wiegand, T. & Wiedenheft, B. Cas9 slide-and-seek for phage defense and genome engineering. *EMBO J* **38**, doi:10.15252/embj.2019101474 (2019).

21 Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62 (2014).

22 Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740 (2009).

23 Shibata, M. *et al.* Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. *Nat Commun* **8**, 1430, doi:10.1038/s41467-017-01466-8 (2017).

24 Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. & Wiedenheft, B. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. *Nucleic Acids Res* **43**, 2216-2222 (2015).

25 Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569 (2014).

26 Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat Commun* **7**, 12778, doi:10.1038/ncomms12778 (2016).

27 Vink, J. N. A. *et al.* Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. *Mol Cell*, doi:10.1016/j.molcel.2019.10.021 (2019).

28 Jackson, R. N., van Erp, P. B., Sternberg, S. H. & Wiedenheft, B. Conformational regulation of CRISPR-associated nucleases. *Current opinion in microbiology* **37**, 110-119 (2017).

29 van Erp, P. B. G. *et al.* Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. *ACS Chem Biol* **13**, 481-490, doi:10.1021/acschembio.7b00649 (2018).

30 Rollins, M. F. *et al.* Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell* **74**, 132-142 e135, doi:10.1016/j.molcel.2019.02.001 (2019).

31 Yang, M. *et al.* The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. *Cell Rep* **22**, 372-382, doi:10.1016/j.celrep.2017.12.048 (2018).

32 Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, doi:10.1126/science.aat0839 (2018).

33 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014).

34 Sun, W. *et al.* Structures of Neisseria meningitidis Cas9 Complexes in Catalytically Poised and Anti-CRISPR-Inhibited States. *Mol Cell*, doi:10.1016/j.molcel.2019.09.025 (2019).

35 Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science advances* **3**, eaao0027 (2017).

36 Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science* **357**, 1420-1424 (2017).

37 Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* **190**, 1390-1400, doi:10.1128/JB.01412-07 (2008).

38 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* **108**, 10098-10103, doi:10.1073/pnas.1104144108 (2011).

39 Bryson, A. L. *et al.* Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. *MBio* **6**, e00648, doi:10.1128/mBio.00648-15 (2015).

40 Vlot, M. *et al.* Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. *Nucleic acids research* **46**, 873-885 (2017).

41 Borges, A. L. *et al.* Bacteriophage Cooperation Suppresses CRISPR-Cas3 and Cas9 Immunity. *Cell* **174**, 917-925.e910, doi:10.1016/j.cell.2018.06.013 (2018).

42 Landsberger, M. *et al.* Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. *Cell* **174**, 908-916.e912, doi:10.1016/j.cell.2018.05.058 (2018).

43 Stone, N. P. *et al.* A hyperthermophilic phage decoration protein suggests common evolutionary origin with herpesvirus triplex proteins and an anti-CRISPR protein. *Structure* **26**, 936-947. e933 (2018).

44 Birkholz, N., Fagerlund, R. D., Smith, L. M., Jackson, S. A. & Fineran, P. C. The autoregulator Aca2 mediates anti-CRISPR repression. *Nucleic Acids Res* **47**, 9658-9665, doi:10.1093/nar/gkz721 (2019).

45 Stanley, S. Y. *et al.* Anti-CRISPR-Associated Proteins Are Crucial Repressors of Anti-CRISPR Transcription. *Cell* **178**, 1452-1464. e1413 (2019).

46 Bondy-Denomy, J. *et al.* A unified resource for tracking anti-CRISPR names. *The CRISPR journal* **1**, 304-305 (2018).

47 Pawluk, A. *et al.* Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat Microbiol* **1**, 16085, doi:10.1038/nmicrobiol.2016.85 (2016).

48 Vorontsova, D. *et al.* Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res* **43**, 10848-10860, doi:10.1093/nar/gkv1261 (2015).

49 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202, doi:10.1038/nature14245 (2015).

50 Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-361, doi:10.1101/gad.257550.114 (2015).

51 Hurford, A. & Day, T. Immune evasion and the evolution of molecular mimicry in parasites. *Evolution* **67**, 2889-2904, doi:10.1111/evo.12171 (2013).

52 Putnam, C. D. & Tainer, J. A. Protein mimicry of DNA and pathway regulation. *DNA Repair (Amst)* **4**, 1410-1420, doi:10.1016/j.dnarep.2005.08.007 (2005).

53 Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-57 e11, doi:10.1016/j.cell.2017.03.012 (2017).

54 Guo, T. W. *et al.* Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426 e412, doi:10.1016/j.cell.2017.09.006 (2017).

55 Peng, R. *et al.* Alternate binding modes of anti-CRISPR viral suppressors AcrF1/2 to Csy surveillance complex revealed by cryo-EM structures. *Cell Res* **27**, 853-864, doi:10.1038/cr.2017.79 (2017).

56 Liu, L., Yin, M., Wang, M. & Wang, Y. Phage AcrIIA2 DNA Mimicry: Structural Basis of the CRISPR and Anti-CRISPR Arms Race. *Mol Cell* **73**, 611-620 e613, doi:10.1016/j.molcel.2018.11.011 (2019).

57 Jiang, F. *et al.* Temperature-responsive competitive inhibition of CRISPR-Cas9. *Mol Cell* **73**, 601-610. e605 (2019).

58 Yang, H. & Patel, D. J. Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. *Mol Cell* **67**, 117-127 e115, doi:10.1016/j.molcel.2017.05.024 (2017).

59 Shin, J. *et al.* Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci Adv* **3**, e1701620, doi:10.1126/sciadv.1701620 (2017).

60 Dong *et al.* Structural basis of CRISPR-SpyCas9 inhibition by an anti-CRISPR protein. *Nature* **546**, 436-439, doi:10.1038/nature22377 (2017).

61 Wang, J. *et al.* A CRISPR evolutionary arms race: structural insights into viral anti-CRISPR/Cas responses. *Cell Res* **26**, 1165 (2016).

62 Wang, X. *et al.* Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. *Nat Struct Mol Biol* **23**, 868-870, doi:10.1038/nsmb.3269 (2016).

63 Zhu, Y. *et al.* Diverse Mechanisms of CRISPR-Cas9 Inhibition by Type IIC Anti-CRISPR Proteins. *Mol Cell* **74**, 296-309 e297, doi:10.1016/j.molcel.2019.01.038 (2019).

64 Thavalingam, A. *et al.* Inhibition of CRISPR-Cas9 ribonucleoprotein complex assembly by anti-CRISPR AcrIIC2. *Nat Commun* **10**, 2806, doi:10.1038/s41467-019-10577-3 (2019).

65 Fuchsbauer, O. *et al.* Cas9 Allosteric Inhibition by the Anti-CRISPR Protein AcrIIA6. *Mol Cell*, doi:10.1016/j.molcel.2019.09.012 (2019).

66 Knott, G. J. *et al.* Structural basis for AcrVA4 inhibition of specific CRISPR-Cas12a. *Elife* **8**, doi:10.7554/eLife.49110 (2019).

67 Peng, R. *et al.* Structural insight into multistage inhibition of CRISPR-Cas12a by AcrVA4. *Proceedings of the National Academy of Sciences* **116**, 18928-18936 (2019).

68 Zhang, H. *et al.* Structural Basis for the Inhibition of CRISPR-Cas12a by Anti-CRISPR Proteins. *Cell Host Microbe* **25**, 815-826 e814, doi:10.1016/j.chom.2019.05.004 (2019).

69 Knott, G. J. *et al.* Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. *Nat Struct Mol Biol* **26**, 315-321, doi:10.1038/s41594-019-0208-z (2019).

70 Ka, D., An, S. Y., Suh, J.-Y. & Bae, E. Crystal structure of an anti-CRISPR protein, AcrIIA1. *Nucleic Acids Res* **46**, 485-492 (2017).

71 Hynes, A. P. *et al.* Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. *Nat Commun* **9**, 2919, doi:10.1038/s41467-018-05092-w (2018).

72 Watters, K. E., Fellmann, C., Bai, H. B., Ren, S. M. & Doudna, J. A. Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science* **362**, 236-239 (2018).

73 Osuna, B. A. *et al.* Listeria phages induce Cas9 degradation to protect lysogenic genomes. *BioRxiv*, 787200 (2019).

74 Rollins, M. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci U S A* **114**, E5113-E5121, doi:10.1073/pnas.1616395114 (2017).

75 Richter, C., Gristwood, T., Clulow, J. S. & Fineran, P. C. In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype IF CRISPR/Cas System. *PloS one* **7**, e49549 (2012).

76 Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci U S A* **114**, E5122-E5128, doi:10.1073/pnas.1618421114 (2017).

77 Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci U S A* **108**, 10092-10097, doi:10.1073/pnas.1102716108 (2011).

78 Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. *J Bacteriol* **194**, 5728-5738 (2012).

79 Kim, Y. *et al.* Anti-CRISPR AcrIIC3 discriminates between Cas9 orthologs via targeting the variable surface of the HNH nuclease domain. *The FEBS journal* (2019).

80 Dong, L. *et al.* An anti-CRISPR protein disables type V Cas12a by acetylation. *Nat Struct Mol Biol* **26**, 308-314, doi:10.1038/s41594-019-0206-1 (2019).

81 Stella, S. *et al.* Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. *Cell* **175**, 1856-1871 e1821, doi:10.1016/j.cell.2018.10.045 (2018).

82 Salah Ud-Din, A. I., Tikhomirova, A. & Roujeinikova, A. Structure and Functional Diversity of GCN5-Related N-Acetyltransferases (GNAT). *Int J Mol Sci* **17**, doi:10.3390/ijms17071018 (2016).

83 Hidalgo-Cantabrana, C., Goh, Y. J., Pan, M., Sanozky-Dawes, R. & Barrangou, R. Genome editing using the endogenous type I CRISPR-Cas system in Lactobacillus crispatus. *Proc Natl Acad Sci U S A* **116**, 15774-15783, doi:10.1073/pnas.1905421116 (2019).

84 Niewoehner, O. *et al.* Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548, doi:10.1038/nature23467 (2017).

85 Moon, S. B., Kim, D. Y., Ko, J. H. & Kim, Y. S. Recent advances in the CRISPR genome editing tool set. *Exp Mol Med* **51**, 1-11, doi:10.1038/s12276-019-0339-7 (2019).

86 Athukoralage, J. S. *et al.* An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* **577**, 572-575, doi:10.1038/s41586-019-1909-5 (2020).

87 Wiegand, T. & Wiedenheft, B. CRISPR Surveillance Turns Transposon Taxi. *CRISPR J* **3**, 10-12, doi:10.1089/crispr.2020.29081.twi (2020).

88 Marino, N. D., Pinilla-Redondo, R., Csorgo, B. & Bondy-Denomy, J. Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nat Methods*, doi:10.1038/s41592-020-0771-6 (2020).

89 Bubeck, F. *et al.* Engineered anti-CRISPR proteins for optogenetic control of CRISPR-Cas9. *Nat Methods* **15**, 924-927, doi:10.1038/s41592-018-0178-9 (2018).

90 Basgall, E. M. *et al.* Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and AcrIIA4 in Saccharomyces cerevisiae. *Microbiology* **164**, 464-474, doi:10.1099/mic.0.000635 (2018).

91 Aschenbrenner, S. *et al.* Coupling Cas9 to artificial inhibitory domains enhances CRISPR-Cas9 target specificity. *Sci Adv* **6**, eaay0187, doi:10.1126/sciadv.aay0187 (2020).

92 Dolan, A. E. *et al.* Introducing a Spectrum of Long-Range Genomic Deletions in Human Embryonic Stem Cells Using Type I CRISPR-Cas. *Mol Cell* **74**, 936-950 e935, doi:10.1016/j.molcel.2019.03.014 (2019).

93 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225, doi:10.1038/s41586-019-1323-z (2019).

94 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).

95 Bhoobalan-Chitty, Y., Johansen, T. B., Di Cianni, N. & Peng, X. Inhibition of Type III CRISPR-Cas Immunity by an Archaeal Virus-Encoded Anti-CRISPR Protein. *Cell* **179**, 448-458. e411 (2019).

96 He, F. *et al.* Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype ID immunity. *Nat Microbiol* **3**, 461 (2018).

97 Elde, N. C. *et al.* Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell* **150**, 831-841, doi:10.1016/j.cell.2012.05.049 (2012).

98 Eaglesham, J. B., Pan, Y., Kupper, T. S. & Kranzusch, P. J. Viral and metazoan poxins are cGAMP-specific nucleases that restrict cGAS-STING signalling. *Nature* **566**, 259-263, doi:10.1038/s41586-019-0928-6 (2019).

CHAPTER FOUR

CRISPR SURVEILLANCE TURNS TRANSPOSON TAXI

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Tanner Wiegand

Co-Author: Blake Wiedenheft

Contributions: wrote the manuscript

Manuscript Information

Tanner Wiegand and Blake Wiedenheft

The CRISPR Journal

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Mary Ann Liebert, Inc.
Volume 3
Issue 1
17 February 2020
DOI: 10.1089/crispr.2020.29081.twi

CRISPR SURVEILLANCE TURNS TRANSPOSON TAXI

Tanner Wiegand[1] and Blake Wiedenheft[1]

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

Four Reports Capture Structures of TniQ in Association with a crRNA-guided Surveillance Complex

Like so many "CRISPR stories", this one starts with an observation by Eugene Koonin and collaborators. In this particular instance, Peters *et al* identified variants of the type I CRISPR system (i.e., I-B and I-F) that lack adaptation machinery (*cas1-2*) and the nuclease (*cas3*) necessary for target degradation (Figure 1**)**.[1] It's not uncommon to find degenerate CRISPR systems, so an observation like this might not arouse most of us, but Peters *et al* doubled down. Using comparative genomics and phylogenetics, they showed that these minimal CRISPR-Cas variants associate with distinct families of Tn7-like transposable elements.

Unlike most Tn7 transposons that include *tnsABC*, *tniQ,* and *tnsE*, these particular Tn7-like elements lack the *tnsE* gene, which is responsible for transposition into plasmid DNA. In place of *tnsE*, these Tn7-like transposons encode *cas* genes that are characteristic of a type I CRISPR RNA (crRNA)-guided surveillance complex (Figure 1).[2] Intriguingly, the spacers associated with the CRISPRs of these truncated systems most frequently match sequences derived from plasmids and other mobile genetic elements. This suggests that the crRNA-guided surveillance complex functionally complements TnsE, though it is not clear if or how this system provides an advantage over TnsE-mediated transposition. It's possible that crRNA-guided insertion of the transposon may inhibit replication of the incoming mobile genetic element, which could provide a defensive function for the host, but may limit the spread of the transposon.

In short, it is unclear whether the transposon domesticated the CRISPR, the CRISPR domesticated the transposon, or if the partnership is somehow mutually beneficial.

The Peters *et al* paper finishes with a provocative conclusion, “transposon encoded CRISPR-Cas systems could be harnessed for genome-engineering applications”. This statement resonated with Sam Sternberg’s group. Working with a Tn7-like transposon encoding a variant I-F system, Sternberg’s group showed that the crRNA-guided surveillance complex targets DNA transposition 47–51 bps downstream of the crRNA-guide.[3] Showcasing its versatility, they engineered this system to deliver various genetic payloads to 16 different chromosomal locations with 95% or better on-target integration.

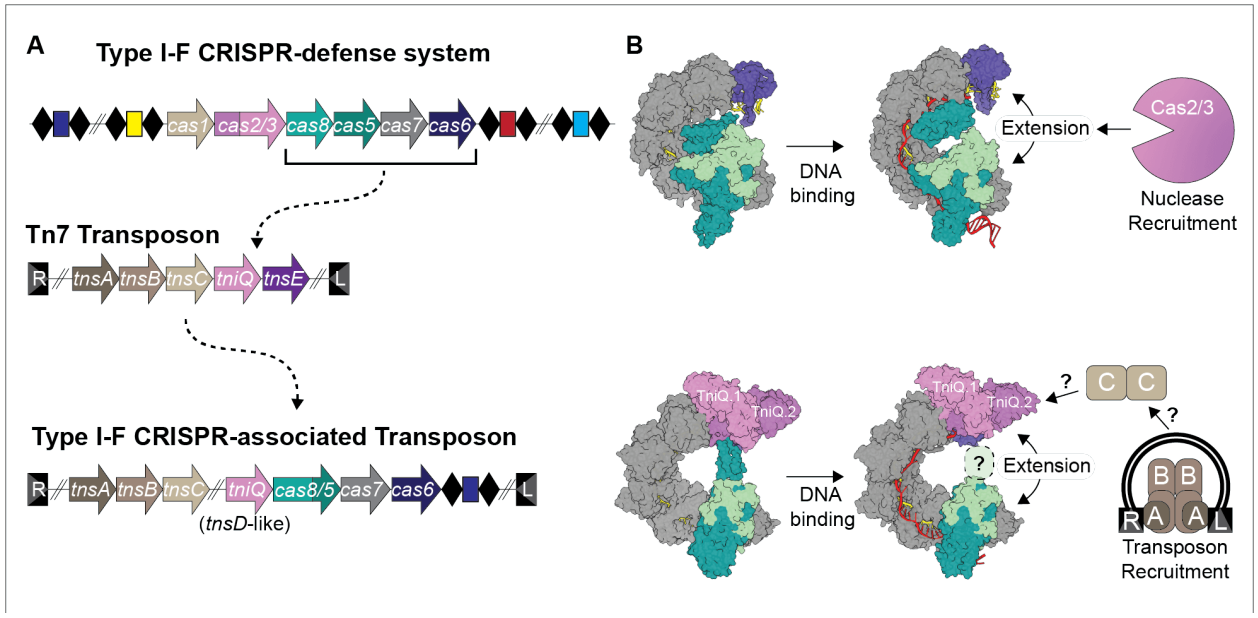


Figure 4.1. Comparison of the I-F CRISPR defense system to the I-F Tn7-like system. (A) Genetic organization of type I-F CRISPR-Cas locus from Pseudomonas aeruginosa PA14 (top). Transposition genes in Tn7 transposon from Escherichia coli (middle). Genetic organization of Tn7-like I-F CRISPR-Cas locus from Vibrio cholerae strain HE-45 (bottom). (B) Structures of the type I-F surveillance complex before (PDB: 5UZ9) and after (PDB: 6NE0) binding dsDNA. The R-loop induced conformational change reveals a recruitment signal for Cas2-3 transacting nuclease (top). TniQ-VcCascade undergoes conformational changes on target hybridization (PDB: 6UVN, 6PIJ; bottom). The mechanism of recruitment remains unknown

Hell Breaks Loose

Eric Sontheimer was recently quoted as saying, “Like everything else in the CRISPR world, the thin edge of the wedge comes in, and the next thing you know all hell breaks loose.”[4] Sontheimer’s statement was in reference to the use of anti-CRISPRs for applications in genome engineering, but these words apply to most research in this field, including crRNA-guided transposons. In the past few months, four papers have been published on structures of transposon-encoded CRISPR–Cas systems. These include Sternberg’s group in collaboration with Israel Fernández in *Nature* last December (5 months after their preprint was posted on bioRxiv) and three other teams writing in *Cell Research*.[5-8]

Collectively, these papers report a total of 10 structures including crystal structures of TniQ, and cryo-EM structures of TniQ in association with the I-F crRNA-guided surveillance complex from *Vibrio cholerae* (i.e., VcCascade-TniQ), both before and after binding to a DNA target. The minimal I-F crRNA-guided surveillance complex maintains a crescent shaped backbone, comprised of six Cas7 subunits, and is capped on either end by a Cas6 “head” and a Cas8/5 “tail” (Fig 1B**)**. This overall shape is similar to what has been observed in structures of I-F surveillance complexes involved in phage defense (Fig 1B).[9-13] Moreover, crRNA-guided DNA binding triggers an elongation in the backbone.[10,12] In the I-F defense systems, elongation of the backbone coincides with a dramatic rearrangement of a helical bundle on Cas8.[12] This rearrangement locks the complex onto the DNA target and presents a landing pad for nuclease recruitment. In the Tn7-like I-F CRISPR system, Cas8 and Cas5 are fused (Cas8/5) and the corresponding helical bundle is positioned like a bridge that connects the Cas6 head to the tail of the complex. Density for this region of the protein is weak and the importance of this position

remains unclear, but its location suggests that it may be involved in stabilizing the displaced strand of DNA (R-loop) and/or coordinated recruitment of transposon-associated proteins. Contacts between the helical bundle of Cas8/5 and TniQ, which forms a homodimer that is prominently positioned at the head of the complex, offer additional support for a role in coordinated recruitment of transposon-associated proteins (Fig 1B).

While this work has revealed specific interactions between VcCasacade and TniQ, additional structures that include TnsABC will help clarify the mechanism of transposon recruitment. We anticipate that additional mechanistic insights will help develop these systems for creative applications in genome engineering. Once these, and other transposon associated CRISPR systems[14] are used to deliver a therapeutic payload to a specific genetic address in the human genome, then “all hell” is going to break loose.

Acknowledgements

We thank Halpin-Healy et al., Li et al., and Jia et al. for providing atomic coordinates ahead of release by the PDB. B.W. is the founder of SurGene, LLC, and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. Research in the Wiedenheft lab is supported by the National Institutes of Health (R35GM134867) and the Montana State University Agricultural Experimental Station (USDA NIFA).

References

1 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci U S A* **114**, E7358-E7366, doi:10.1073/pnas.1709035114 (2017).

2 Peters, J. E. Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond. *Mol Microbiol* **112**, 1635-1644, doi:10.1111/mmi.14383 (2019).

3 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225, doi:10.1038/s41586-019-1323-z (2019).

4 Dolgin, E. The kill-switch for CRISPR that could make gene-editing safer. *Nature* **577**, 308-310, doi:10.1038/d41586-020-00053-0 (2020).

5 Jia, N., Xie, W., de la Cruz, M. J., Eng, E. T. & Patel, D. J. Structure-function insights into the initial step of DNA integration by a CRISPR-Cas-Transposon complex. *Cell Res*, doi:10.1038/s41422-019-0272-2 (2020).

6 Wang, B., Xu, W. & Yang, H. Structural basis of a Tn7-like transposase recruitment and DNA loading to CRISPR-Cas surveillance complex. *Cell Res*, doi:10.1038/s41422-020-0274-0 (2020).

7 Li, Z., Zhang, H., Xiao, R. & Chang, L. Cryo-EM structure of a type I-F CRISPR RNA guided surveillance complex bound to transposition protein TniQ. *Cell Res*, doi:10.1038/s41422-019-0268-y (2020).

8 Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernandez, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. *Nature* **577**, 271-274, doi:10.1038/s41586-019-1849-0 (2020).

9 Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-57 e11, doi:10.1016/j.cell.2017.03.012 (2017).

10 Guo, T. W. *et al.* Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426 e412, doi:10.1016/j.cell.2017.09.006 (2017).

11 Peng, R. *et al.* Alternate binding modes of anti-CRISPR viral suppressors AcrF1/2 to Csy surveillance complex revealed by cryo-EM structures. *Cell Res* **27**, 853-864, doi:10.1038/cr.2017.79 (2017).

12 Rollins, M. F. *et al.* Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell* **74**, 132-142 e135, doi:10.1016/j.molcel.2019.02.001 (2019).

13 Pausch, P. *et al.* Structural Variation of Type I-F CRISPR RNA Guided DNA Surveillance. *Mol Cell* **67**, 622-632 e624, doi:10.1016/j.molcel.2017.06.036 (2017).

14 Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48-53, doi:10.1126/science.aax9181 (2019).

CHAPTER FIVE

CAS9 SLIDE AND SEEK FOR PHAGE DEFENSE

Contribution of Authors and Co-Authors

Manuscript in Chapter 5

Author: Andrew Santiago-Frangos

Contributions: wrote the manuscript

Co-Author: Tanner Wiegand

Contributions: wrote the manuscript

Co-Author: Blake Wiedenheft

Contributions: wrote the manuscript

Manuscript Information

Andrew Santiago-Frangos, Tanner Wiegand, and Blake Wiedenheft

The EMBO Journal

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

EMBO press
Volume 38
Issue 4
15 February 2019
DOI: 10.15252/embj.2019101474

CAS9 SLIDE AND SEEK FOR PHAGE DEFENSE

Andrew Santiago-Frangos,[1] Tanner Wiegand,[1] and Blake Wiedenheft[1]

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

Abstract

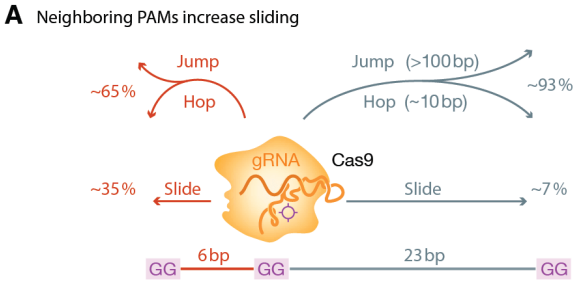
How does the Cas9 nuclease locate a specific 20-nucleotide target sequence in a crowded intracellular environment packed with mega bases of distracting non-target DNA? Previously, it was shown that Cas9 finds DNA targets via three-dimensional diffusion. In this issue of *The EMBO Journal*, Globyte et al (2019) reveal another dimension of the search process, which involves short-range one-dimensional sliding. These results have implications for understanding the natural function of Cas9 and its applications in genome engineering experiments.

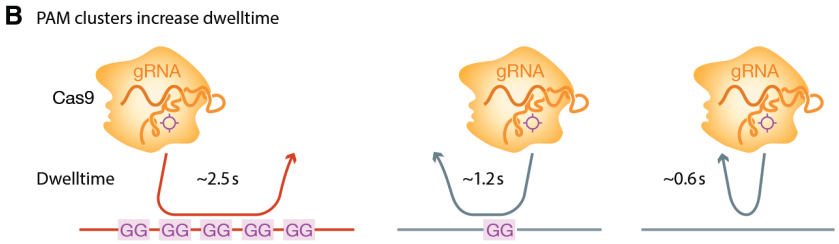
Main Text

Cas9 is an RNA-guided nuclease that is routinely used for targeted genome engineering in eukaryotic cells. However, the roles and responsibilities of genome engineers are profoundly different from the natural selective pressures that optimized Cas9 for protection from invading genetic parasites, like phages. To optimize Cas9 for applications in genome engineering, it is critical to understand the biophysical mechanisms that govern target identification. In this issue of *The EMBO Journal*, Globyte et al (2019) show that the Cas9 target search process involves a combination of 1D sliding and 3D diffusion.[1]

Unwinding all the double-stranded DNA (dsDNA) in a cell to find a complementary target would be a slow and energetically expensive process. Instead, target identification by Cas9 first relies on detection of a short, double-stranded sequence motif known as the PAM (protospacer adjacent motif).[2] PAM recognition by Cas9 destabilizes the duplex and facilitates RNA-guided strand invasion for detection of a complementary (protospacer) target.[3] This sequential process facilitates target finding in a crowded intracellular environment. But how does Cas9 locate PAMs, how long does it take to find the PAM with an adjacent complementary (protospacer) target, and how does the landscape of this search change when we move Cas9 from a small bacterial cell containing ~600 lm of DNA to a human nucleus containing over 2 m of DNA (Fig 5.1)?

Many DNA-binding proteins, such as the Lac repressor LacI, find their targets at speeds that approach, or surpass, the limits of 3-dimensional diffusion.[4] These proteins commonly rely on a combination of 1D sliding and 3D hopping or jumping. Sliding involves short-range movements along the contour of the DNA, whereas hopping and jumping require dissociation





C

	Length of genomic DNA per cell or phage	Number of total PAMs in genome	Percentage of PAMs within ±25 bp of another PAM	Percentage of nucleotides in genome within ±25 bp of a PAM
Homo sapiens	2,100,000 µm	>322,340,476	~94.12%	~83.58%
Streptococcus pyogenes	630 µm	142,700	86.33%	82.42%
A25 bacteriophage	11.5 µm	2,614	86.99%	89.68%

© EMBO

Figure 5.1. PAM-mediated dwell times of Cas9. (A) Schematic representation of Cas9 one-dimensional diffusion (sliding) with probabilities assigned by Globyte et al (2019) (B) Schematic of Cas9 average dwell times (Δτav) reported by Globyte et al (2019) on DNA substrates with different number of PAMs. (C) Physical lengths and PAMs distributions in genomes of Homo sapiens, Streptococcus pyogenes, and bacteriophage A25.

and rebinding events over short (~10-bps, hopping) or long distances, respectively. The relative contributions of sliding, hopping, and jumping have been optimized by evolution for each target-protein pair.[4]

Single-molecule tracking of Cas9 has previously shown that Cas9 predominantly uses 3D diffusion (e.g., jumping) to locate DNA targets.[3,5,6] Cas9 frequently samples off-target DNA, where it dwells for an average of 750 ms in human cells or 30 ms in *Escherichia coli*.[5,6] However, previous methods did not have the resolution to detect short-range sliding, representing a gap in our knowledge of how Cas9 locates targets.

In this issue of *The EMBO Journal*, Globyte *et al* (2019) address this knowledge gap using single-molecule Förster Resonance Energy Transfer (smFRET) assays designed to examine the behavior of Cas9 with high spatial and temporal resolution.[1] Their results suggest that Cas9 frequently slides between adjacent PAMs, and that the probability of sliding to a neighboring PAM decreases exponentially as the distance traveled increases (Fig 5.1). Models of target search by a DNA-binding protein, which use sliding, hopping, and jumping to find its target rapidly, suggest the optimal sliding length is ~10 times the size of its target. Therefore, the max sliding distance measured (i.e., ~25 base pairs) suggests that sliding by Cas9 has been optimized for rapid detection of short PAMs, rather than the longer protospacer.

Additionally, Globyte *et al* show that the average dwell time of Cas9 is nearly 2-fold longer (2.5 s) on DNA containing 5 PAMs in close proximity, compared to a sequence containing 1 PAM (Fig 5.1).[1] They show that this increased dwell time is due to shuttling of Cas9 between nearby PAMs. These observations help explain why Cas9 bound more strongly to PAM-rich regions of DNA,[3] and illustrate how PAM-rich zones may sequester Cas9 in vivo. 86% of the 142,700 PAMs in the *Streptococcus pyogenes* genome lie within 25 bps of another PAM, which suggests that they will frequently be sampled by sliding. The human genome contains three orders of magnitude more PAMs, and an even larger proportion (i.e., 94%) of

these are within 25 bps of another PAM. Therefore, Cas9 variants that rely heavily on sliding during target search are likely to be sequestered the many PAM-rich islands present within the human genome.

In the natural context of a bacterial cell, and in the absence of an infection, a certain population of Cas9 may associate with a few high affinity, PAM-rich islands on the host genome, while the remaining Cas9s are free to search for PAMs and complementary DNA targets. When foreign DNA is introduced to the cell, Cas9 associates with new DNA via long-range jumping. However, once Cas9 has bound to this new segment of DNA, sliding and hopping behaviors are important for ensuring that Cas9 interrogates the foreign DNA. The Cas9 slide-and-seek behavior described by Globyte *et al* (2019) is an important new dimension to the target search process that protects bacterial cells from infection.[1]

Recently, many strategies have been developed to redesign Cas9 effectors and the RNA guide for improved gene editing specificity.[7] Mutations of *S. pyogenes* Cas9 that remove non-sequence specific interactions with the phosphate backbone of the complementary DNA strand (SpCas9-HF1) significantly reduced off-target cleavage events and DNA association rates, while retaining most on-target activity.[8,9] Therefore, it is possible that mutations in SpCas9-HF1 decreased the relative contributions of sliding and hopping in favor of jumping. Thus, mutational manipulation of search mechanisms may re-optimize Cas9 for target searching in the context of a human genome, where a jump-and-seek approach may be preferable to slide-and-seek.

References

1 Globyte, V., Lee, S. H., Bae, T., Kim, J. S. & Joo, C. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *EMBO J* **38**, doi:10.15252/embj.201899466 (2019).

2 Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569 (2014).

3 Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62 (2014).

4 Halford, S. E. An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem Soc Trans* **37**, 343-348, doi:10.1042/BST0370343 (2009).

5 Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science* **357**, 1420-1424 (2017).

6 Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823-826, doi:10.1126/science.aac6572 (2015).

7 Wilkinson, R. A., Martin, C., Nemudryi, A. A. & Wiedenheft, B. CRISPR RNA-guided autonomous delivery of Cas9. *Nature structural & molecular biology* **26**, 14-24 (2019).

8 Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495, doi:10.1038/nature16526 (2016).

9 Singh, D. *et al.* Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. *Nat Struct Mol Biol* **25**, 347-354, doi:10.1038/s41594-018-0051-7 (2018).

CHAPTER SIX

TEMPORAL DETECTION AND PHYLOGENETIC ASSESSMENT OF SARS-COV-2 IN MUNICIPAL WASTEWATER

Contribution of Authors and Co-Authors

Manuscript in Chapter 6

Author: Artem Nemudryi

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript, wrote the manuscript

Co-Author: Anna Nemudraia

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Tanner Wiegand

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Kevin Surya

Contributions: performed the research

Co-Author: Murat Buyukyoruk

Contributions: performed the research

Co-Author: Calvin Cicha

Contributions: performed the research

Co-Author: Karl K. Vanderwood

Contributions: performed the research

Co-Author: Royce Wilkonson

Contributions: performed the research

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

Artem Nemudryi, Anna Nemudraia, Tanner Wiegand, Kevin Surya Murat Buyukyoruk, Calvin Cicha, Karl K. Vanderwood, Royce Wilkinson, and Blake Wiedenheft

Cell Reports Medicine

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Cell Press
Volume 1
Issue 6
22 September 2020
DOI: 10.1016/j.xcrm.2020.100098

TEMPORAL DETECTION AND PHYLOGENETIC ASSESSMENT OF SARS-COV-2 IN MUNICIPAL WASTEWATER

Artem Nemudryi,[1] Anna Nemudraia,[1] Tanner Wiegand,[1] Kevin Surya,[1] Murat Buyukyoruk,[1] Calvin Cicha,[1] Karl K. Vanderwood,[2] Royce Wilkinson,[1] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Gallatin City-County Health Department, Bozeman, MT 59715, USA.

Abstract

SARS-CoV-2 has recently been detected in feces, which indicates that wastewater may be used to monitor viral prevalence in the community. Here, we use RT-qPCR to monitor wastewater for SARS-CoV-2 RNA over a 74-day time course. We show that changes in SARS-CoV-2 RNA concentrations follow symptom onset gathered by retrospective interview of patients but precedes clinical test results. In addition, we determine a nearly complete (98.5%) SARS-CoV-2 genome sequence from wastewater and use phylogenetic analysis to infer viral ancestry. Collectively, this work demonstrates how wastewater can be used as a proxy to monitor viral prevalence in the community and how genome sequencing can be used for genotyping viral strains circulating in a community.

Introduction

In late December 2019, authorities from the People's Republic of China (PRC) announced an epidemic of pneumonia.[1] A novel coronavirus (severe acute respiratory syndrome-coronavirus-2 [SARS-CoV-2]) was identified as the etiologic agent and the disease was named coronavirus disease 2019 (COVID-19). The virus spread rapidly, first to Thailand, Japan, Korea, and Europe, and now to >188 countries across all of the continents except Antarctica. The global total of infected individuals now exceeds 20 million (https://coronavirus.jhu.edu/).

Public health professionals around the world are working to limit the spread of SARS-CoV-2, and ''flatten the curve,'' which requires a reduction in cases from one day to the next. However, SARS-CoV-2 containment has been outpaced by viral spread and limited resources for testing. Moreover, mounting evidence suggests that the virus is not only spread by aerosols but may also be transmitted via feces. Both viral RNA and infectious virus have been detected in the stool of COVID-19 patients.[2-8] This has important implications for the spread of the virus and suggests that wastewater may be used to monitor progression or abatement of viral spread at the community level[9-17]

Results

To test whether wastewater could be used for SARS-CoV-2 surveillance, we collected samples from the municipal wastewater treatment plant in Bozeman, Montana (USA). Untreated wastewater samples were collected on 17 different days over the course of a 74-day period using an autosampler that collects a volume proportional to flow for 24 h (Table S6.1). This composite sample reflects the average characteristics of wastewater over the previous day. The samples

were filtered and concentrated before RNA extraction. To concentrate SARS-CoV-2, we used ultrafiltration with spin concentrators that efficiently recover viruses from wastewater.[11,18] Extracted RNA was used as a template for one-step quantitative reverse-transcriptase-polymerase chain reaction (qRT-PCR), performed according to Centers for Disease Control and Prevention (CDC) guidelines (https://www.fda.gov/media/134922/ download). Each qRT-PCR reaction was performed using 2 primer pairs (N1 and N2), which target distinct regions of the nucleocapsid (N) gene from SARS-CoV-2 (Figures 6.1A and S6.1). Composite samples collected in late March and early April 2020 tested positive for SARS-CoV-2, although concentrations of viral RNA steadily declined and then dropped below the limit of detection. After 1 month of undetectable levels of SARS-CoV-2, the wastewater began testing positive again in late May, which coincided with an increase in COVID-19 cases in the community (Figure 6.1A).

The current methods for tracking the COVID-19 pandemic primarily rely on clinical test results, but this process involves intrinsic delays that preclude real-time tracking of the outbreak. On average, a person develops symptoms 4–5 days after initial exposure, and it is predicted that only 32% of symptomatic individuals are tested.[19,20] Test results are typically available 3–9 days after illness onset.[14,21] We hypothesized that wastewater levels of SARS-CoV-2 RNA correlate with COVID-19 incidence rates and that these data could be used as an epidemiological indicator to track the outbreak in real time. To test this hypothesis, we compared our wastewater surveillance data to the frequencies of reported lab-confirmed cases and symptom onset dates that were collected by retrospective interviews. In the initial outbreak (mid-March 2020), the SARS-CoV-2 RNA concentration in the wastewater lagged behind symptom onset data by 8

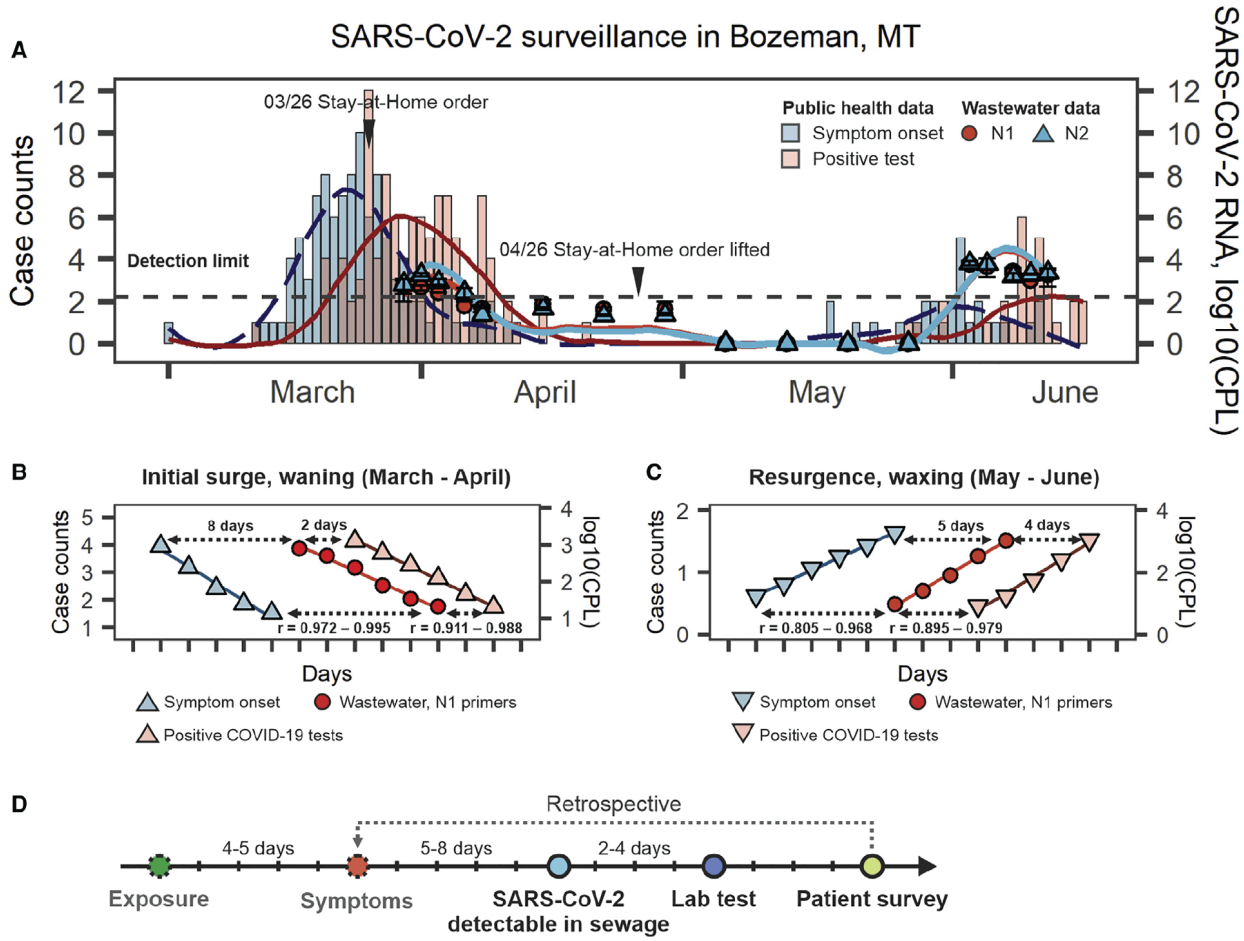


Figure 6.1. Detection and Quantification of SARS-CoV-2 in Wastewater and in the Community (A) Temporal dynamics of SARS-CoV-2 RNA in the municipal wastewater is superimposed on the epidemiological data. Symptom onset data (teal bars) were collected by retrospective interviews of COVID-19 patients who previously tested positive for SARS-CoV-2 (coral bars). The red circles and blue triangles show SARS-CoV-2 RNA concentration in municipal wastewater (means ± SDs) measured with qRT-PCR using the N1 and N2 primer pairs, respectively (see Method Details). The lines show curves fitted to qRT-PCR and epidemiological data using local polynomial regression (LOESS, locally estimated scatterplot smoothing). (B and C) Linear regions of the epidemiological and wastewater curves. Curves were displaced relative to each other and Pearson correlation coefficients (r) were calculated. The 95% confidence intervals for the highest Pearson's r values and respective offsets are shown. Data for initial surge (March–April) and resurgence (May) were analyzed separately. Surge boundaries were defined as the earliest reported symptom onset (left boundary) and date with last reported positive test (right boundary). The interval between surges with zero reported cases/symptoms (mid-April–mid-May) was dropped from the analysis. (D) Timeline of the indicators used in the study. Symptom onset is the earliest available estimate of the viral spread. However, these data are collected retrospectively, which preclude its use for real-time tracking of the outbreak. Wastewater correlates with symptom onset and could be used to track progressing outbreak.

days (Pearson's r = 0.989; Figure 6.1B), and preceded laboratory test results for individuals by 2 days (r = 0.969; Figure 6.1B). When cases resurged in May, wastewater detection trailed symptom onset by 5 days (r = 0.92) and foreshadowed the increase in positive tests by 4 days (r = 0.953; Figure 6.1C).While wastewater detection trailed symptom onset by 5 days, it is important to note that the retrospective interviews used to collect symptom onset information are available only ~10 days after exposure. Our analysis demonstrates that wastewater surveillance is the earliest real-time measure of SARS-CoV-2 prevalence (Figure 6.1D).

To verify that the qRT-PCR results reflect bona fide detection of SARS-CoV-2 rather than priming from an unintended template, we repeated the PCR using 10 primer pairs that tile across the SARS-CoV-2 genome.[22] These primers were designed to target conserved regions of the genome that flank polymorphic sites that have been used to trace viral ancestry and geographic origins (Figure S6.2A).[22,23] RNA isolated from the Bozeman waste stream on March 27, 2020 was used as a template for these RT-PCR reactions, and all 10 primer pairs produced PCR products of the expected sizes (Figure S6.2B; Table S6.2). PCR products were sequenced using the Sanger method and the reads were aligned to the reference genome using MUSCLE.[24,25] We observed no sequence heterogeneity in redundant reads derived from each location of the genome (Figure S6.2C). The same RNA sample was further used to determine a nearly complete (98.5%) SARS-CoV-2 genome sequence using a long-read sequencing platform.[22,26]

Mutations that do not confer a fitness defect are preserved in viral progeny, and thus serve as genetic landmarks that can be used to trace viral ancestry. Efforts to understand the origins and evolution of SARS-CoV-2 have resulted in ~82,000 genome sequences from 91

countries as of August 13, 2020 (https:// www.gisaid.org/). Phylogenetic analyses of these sequences have enabled molecular tracking of viral spread.[27-32] To determine the ancestry of SARS-CoV-2 strains circulating in Bozeman's wastewater on March 27, we determined the genome sequence using Oxford Nanopore. Approximately 2,000 copies of the viral RNA (estimated with qPCR) were used to generate an amplicon library.[26,33] Nanopore sequencing on the MinION platform resulted in ~700,000 reads. Quality control and base calling were performed with MinKNOW version 19.06.8 in High Accuracy mode (Oxford Nanopore Technologies), and the sequences were assembled using the bioinformatic pipeline from ARTIC Network (https://artic.network/ncov-2019). This approach resulted in a single viral contig with an average sequencing depth of 6,875x that covered 98.5% of the SARS-CoV-2 reference genome (GenBank: MN908947.3). Unsequenced regions of the genome include the 5' and 3' ends and a stretch of 170 bases (22,346–22,515 in the Wuhan-Hu-1 reference genome), which likely had too few reads for basecalling due to PCR bias.[34]

In total, we found 11 single-nucleotide variants (SNVs) in the assembled genome that distinguish the Bozeman wastewater SARS-CoV-2 sequence from the Wuhan-Hu-1/2019 reference sequence (Figure S6.2D). To verify the authenticity of these SNVs, we examined raw sequencing data, which is available on Mendeley (Mendeley Data: https://dx.doi.org/10.17632/nfsfvy6xkf.1). During this analysis, we noticed that one of the variants (A23122T) was introduced by incomplete trimming of a sequencing adaptor (5'-CGTATTGCT) that is partially homologous (underlined) to the reference genome (5'-TACATGCA). This variant has been reported in other genomes sequenced using this protocol,[35] and similar issues have been identified in other regions of the SARS-CoV-2 genome.[36] In contrast to A23122T, we found no

other evidence for trimming artifacts for the 10 remaining SNVs. It is possible that the consensus genome presented here is derived from a chimeric assembly of distinct genotypes, but >90% of the reads contain the same variant at each of these 10 positions.

The genome was aligned to 14,970 SARS-CoV-2 genomes from 74 different countries (Global Initiative on Sharing All Influenza Data, https://www.gisaid.org/). The resulting alignment was used to build a phylogenetic tree (Figure 6.2A), which indicates that the SARS-CoV-2 genome in Bozeman's wastewater is most closely related to genomes from California and

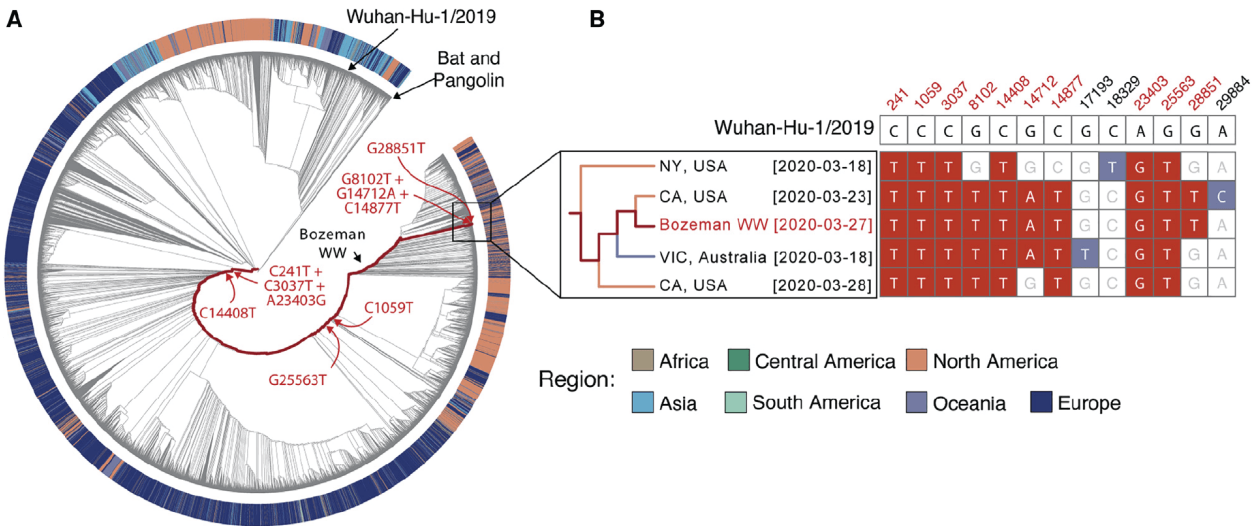


Figure 6.2. Phylogenetic Analysis of SARS-CoV-2 Sequence Isolated from Wastewater. (A) Maximum-likelihood phylogeny of the SARS-CoV-2-related lineage (n = 14,971 sequences). The phylogenetic history of SARS-CoV-2 strain sequenced from Bozeman's wastewater (WW) is shown in crimson. The outer ring is colored according to regions of the world where the samples were isolated. The tree is rooted relative to the RaTG13 genome (a bat coronavirus with 96% sequence similarity to SARS-CoV-2; GenBank: MN996532.1). Mutations that occurred over space and time are shown in red. (B) Sequences isolated from Bozeman WW clade with sequences of US and Australian origin (left). The sequences are named according to the geographic origin and the viral isolation date. A comparison of mutations in sequences is shown in the inset (right). The Wuhan reference sequence for each of the positions where mutations occur is shown across the top. The mutated positions and bases present in Bozeman WW sequence are shown in red, the bases matching Wuhan reference sequence are shown in white, and the mutations not present in the Bozeman WW sequence are shown in blue.

Victoria, Australia). The three mutations that define the Wuhan WA1 linage (C8782T, C18060T, T28144C) are not present in the Bozeman wastewater (WW) genome, while all 10 mutations in the Bozeman wastewater SARS-CoV-2 sequence co-occur in sequences from California; 9 of these 10 mutations are also present in an isolate from Victoria, Australia (Figures 6.2B, S6.2C, and S6.2D).[23,37,38] To determine how these sequence variations may have accumulated over space and time, we mapped each mutation onto the phylogenetic tree of SARS-CoV-2 sequences (Figures 6.2A and S6.2E). This analysis shows that the A28851T mutation has been acquired most recently and confirms that the assembled genome from the Bozeman wastewater is most closely related to a strain circulating in California. While this sequencing approach reveals the genetic history, it does not measure the fitness of this or any of the mutations associated with distinct geographic locations. We anticipate that temporal genome sequencing from the wastewater will help identify viral strains circulating in a specific community over time.

Discussion

The results presented here demonstrate that wastewater monitoring for SARS-CoV-2 RNA by qRT-PCR provides a real-time measure of viral prevalence in the community (Figure 6.1). Clinical testing for COVID-19 typically occurs 3–9 days after symptom onset and may vary, depending on the availability of tests, care-seeking behavior, workloads in testing facilities, and current testing strategy.[21,39,40] In our study, wastewater surveillance for SARS-CoV-2 foreshadowed new case reports by 2–4 days.

The statistics of lab-confirmed COVID-19 cases not only lag behind viral spread but they also underestimate the true scale of the pandemic. A recent analysis of outpatient surveillance data estimates that only 32% of SARS-CoV-2-infected individuals in the United States have

sought medical care.[40] These challenges are overcome with wastewater testing, which captures input from all of the individuals in the local community and thus has the potential for estimating the true prevalence of COVID-19 using computational models that account for the median viral load in stool, virus degradation rates, travel time to the treatment facility, and water use per capita.[14,41] Furthermore, wastewater may capture mild and asymptomatic infections that may be used to alert public health officials about emerging undetected transmission events.[6] Monitoring wastewater for SARS-CoV-2 provides a useful epidemiological metric that could help track the outbreak and inform policy. The study presented here complements the rapidly emerging body of work by providing an important link between wastewater surveillance, COVID-19 epidemiology, and tracing SARS-CoV-2 spread patterns with genome sequencing.

Limitations of Study

Nanopore sequencing has an error rate of ~10%–15%, which precludes the reliable detection of rare genotypes.[33,42] Therefore, the protocol used here is limited to the detection of genotypes with >10%–15% representation. Due to this limitation of the sequencing approach, we cannot exclude the possibility of chimeric genome assembly. Additional sequencing efforts are required to correlate results from wastewater to clinical isolates and determine how ratios of SARS-CoV-2 variants in wastewater translate to ratios in the population.

An additional limitation of this study is the use of a single method for the concentration of SARS-CoV-2 from wastewater. While others have shown that centrifugal ultrafiltration is effective, comparative analyses of alternative concentration protocols, RNA extraction methods, and diagnostic primer sets are required to correlate results from different groups and build a standard approach for wastewater-based SARS-CoV-2 surveillance[9-18].

Acknowledgements

Research in the Wiedenheft lab is supported by the National Institutes of Health (1R35GM134867), the Montana State University (MSU) Agricultural Experimental Station, the M.J. Murdock Charitable Trust, the Gianforte Family Foundation, and the MSU Office of the Vice President for Research. Wastewater data were collected as part of a SARS-CoV-2 surveillance effort supported by the City of Bozeman. We are grateful to Josh French, Justin Roberts, and the other dedicated wastewater technicians who made this work possible. We thank the GISAID EpiFluDatabase and depositing laboratories. The phylogenetic analysis in this article would not have been possible without their willingness to share data. We thank the reviewers for helping identify a mutation that was introduced by incomplete trimming of an adaptor.

References

1 Organization, W. H. (2020).

2 Lodder, W. & de Roda Husman, A. M. SARS-CoV-2 in wastewater: potential health risk, but also data source. *Lancet Gastroenterol Hepatol* **5**, 533-534, doi:10.1016/S2468-1253(20)30087-X (2020).

3 Hindson, J. COVID-19: faecal-oral transmission? *Nat Rev Gastroenterol Hepatol* **17**, 259-259, doi:10.1038/s41575-020-0295-7 (2020).

4 Xu, Y. *et al.* Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature medicine* **26**, 502-505, doi:10.1038/s41591-020-0817-4 (2020).

5 Wu, Y. *et al.* Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *Lancet Gastroenterol Hepatol* **5**, 434-435, doi:10.1016/S2468-1253(20)30083-2 (2020).

6 Tang, A. *et al.* Detection of Novel Coronavirus by RT-PCR in Stool Specimen from Asymptomatic Child, China. *Emerg Infect Dis* **26**, 1337-1339, doi:10.3201/eid2606.200301 (2020).

7 Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA* **323**, 1843-1844, doi:10.1001/jama.2020.3786 (2020).

8 Xiao, F. *et al.* Infectious SARS-CoV-2 in Feces of Patient with Severe COVID-19. *Emerg Infect Dis* **26**, 1920-1922, doi:10.3201/eid2608.200681 (2020).

9 Mallapaty, S. How sewage could reveal true scale of coronavirus outbreak. *Nature* **580**, 176-177, doi:10.1038/d41586-020-00973-x (2020).

10 Randazzo, W. *et al.* SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Res* **181**, 115942-115942, doi:10.1016/j.watres.2020.115942 (2020).

11 Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environmental Science & Technology Letters* **7**, 511-516, doi:10.1021/acs.estlett.0c00357 (2020).

12 Wurtzer, S. *et al.* Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, Greater Paris, France, 5 March to 23 April 2020. *Euro Surveill* **25**, doi:10.2807/1560-7917.ES.2020.25.50.2000776 (2020).

13 Wu, F. *et al.* SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. *mSystems* **5**, doi:10.1128/mSystems.00614-20 (2020).

14 Ahmed, W. *et al.* First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci Total Environ* **728**, 138764-138764, doi:10.1016/j.scitotenv.2020.138764 (2020).

15 Rimoldi, S. G. *et al.* Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. *Sci Total Environ* **744**, 140911-140911, doi:10.1016/j.scitotenv.2020.140911 (2020).

16 La Rosa, G. *et al.* First detection of SARS-CoV-2 in untreated wastewaters in Italy. *Sci Total Environ* **736**, 139652-139652, doi:10.1016/j.scitotenv.2020.139652 (2020).

17 Peccia, J. *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat Biotechnol* **38**, 1164-1167, doi:10.1038/s41587-020-0684-z (2020).

18 Ye, Y., Ellenberg, R. M., Graham, K. E. & Wigginton, K. R. Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater. *Environ Sci Technol* **50**, 5077-5085, doi:10.1021/acs.est.6b00876 (2016).

19 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **382**, 1199-1207, doi:10.1056/NEJMoa2001316 (2020).

20 Lauer, S. A. *et al.* The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Ann Intern Med* **172**, 577-582, doi:10.7326/M20-0504 (2020).

21 Garg, S. *et al.* Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. *MMWR Morb Mortal Wkly Rep* **69**, 458-464, doi:10.15585/mmwr.mm6915e3 (2020).

22 Quick, J. in *protocols.io* (ZappyLab, Inc., 2020).

23 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571-575, doi:10.1126/science.abc0523 (2020).

24 Edgar, R. C. in *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.* (IEEE).

25 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

26 Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature protocols* **12**, 1261-1276, doi:10.1038/nprot.2017.066 (2017).

27 van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* **83**, 104351-104351, doi:10.1016/j.meegid.2020.104351 (2020).

28 Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science (New York, N.Y.)* **369**, 297-301, doi:10.1126/science.abc1917 (2020).

29 Castillo, A. E. *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J Med Virol* **92**, 1562-1566, doi:10.1002/jmv.25797 (2020).

30 Stefanelli, P. *et al.* Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **25**, 2000305, doi:10.2807/1560-7917.ES.2020.25.13.2000305 (2020).

31 Zehender, G. *et al.* Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J Med Virol* **92**, 1637-1640, doi:10.1002/jmv.25794 (2020).

32 Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e995, doi:10.1016/j.cell.2020.04.021 (2020).

33 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome biology* **20**, 8-8, doi:10.1186/s13059-018-1618-7 (2019).

34 Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PloS one* **15**, e0239403-e0239403, doi:10.1371/journal.pone.0239403 (2020).

35 Lopez-Alvarez, D., Parra, B. & Cuellar, W. J. Genome Sequence of SARS-CoV-2 Isolate Cali-01, from Colombia, Obtained Using Oxford Nanopore MinION Sequencing. *Microbiol Resour Announc* **9**, e00573-00520, doi:10.1128/MRA.00573-20 (2020).

36 Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e819, doi:10.1016/j.cell.2020.06.043 (2020).

37 Holshue, M. L. *et al.* First Case of 2019 Novel Coronavirus in the United States. *N Engl J Med* **382**, 929-936, doi:10.1056/NEJMoa2001191 (2020).

38 Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science (New York, N.Y.)* **369**, 582-587, doi:10.1126/science.abb9263 (2020).

39 Biggerstaff, M. *et al.* Influenza-like illness, the time to seek healthcare, and influenza antiviral receipt during the 2010-2011 influenza season-United States. *J Infect Dis* **210**, 535-544, doi:10.1093/infdis/jiu224 (2014).

40 Silverman, J. D., Hupert, N. & Washburne, A. D. Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. *Sci Transl Med* **12**, eabc1126, doi:10.1126/scitranslmed.abc1126 (2020).

41 Hart, O. E. & Halden, R. U. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. *Sci Total Environ* **730**, 138875-138875, doi:10.1016/j.scitotenv.2020.138875 (2020).

42 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338-345, doi:10.1038/nbt.4060 (2018).

43 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

44 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, doi:10.1038/nature16996 (2016).

45 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518, doi:10.1093/nar/gki198 (2005).

46 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

47 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

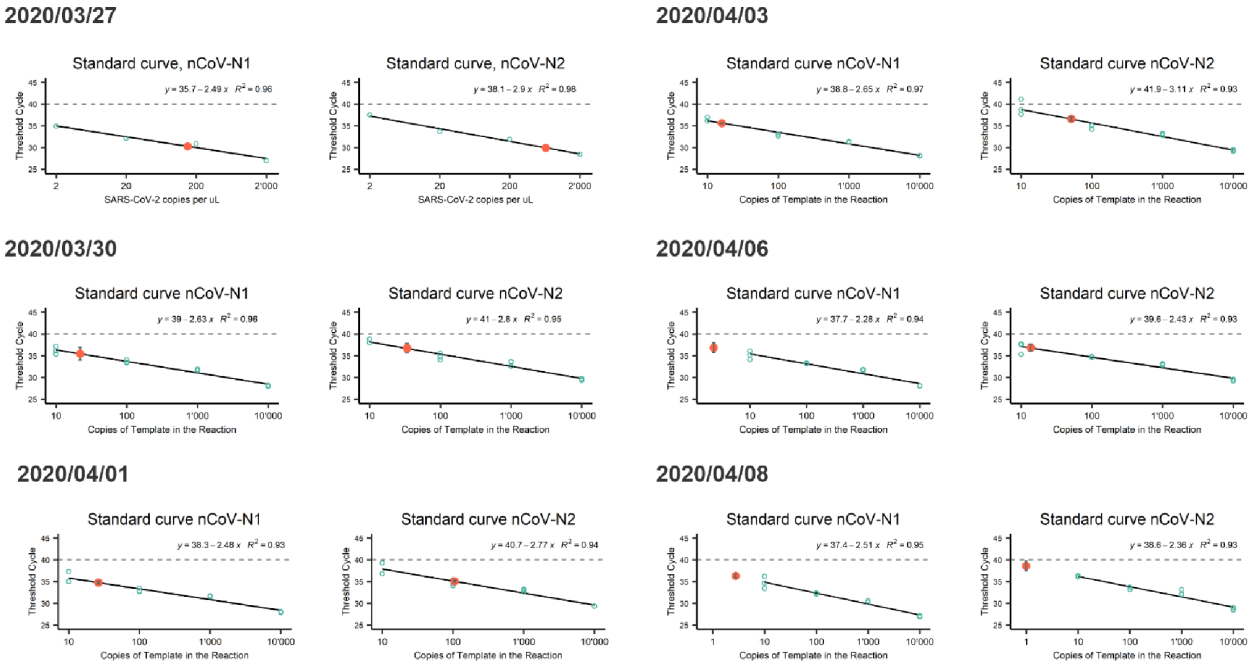
48 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

49 Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28-36, doi:10.1111/2041-210x.12628 (2016).

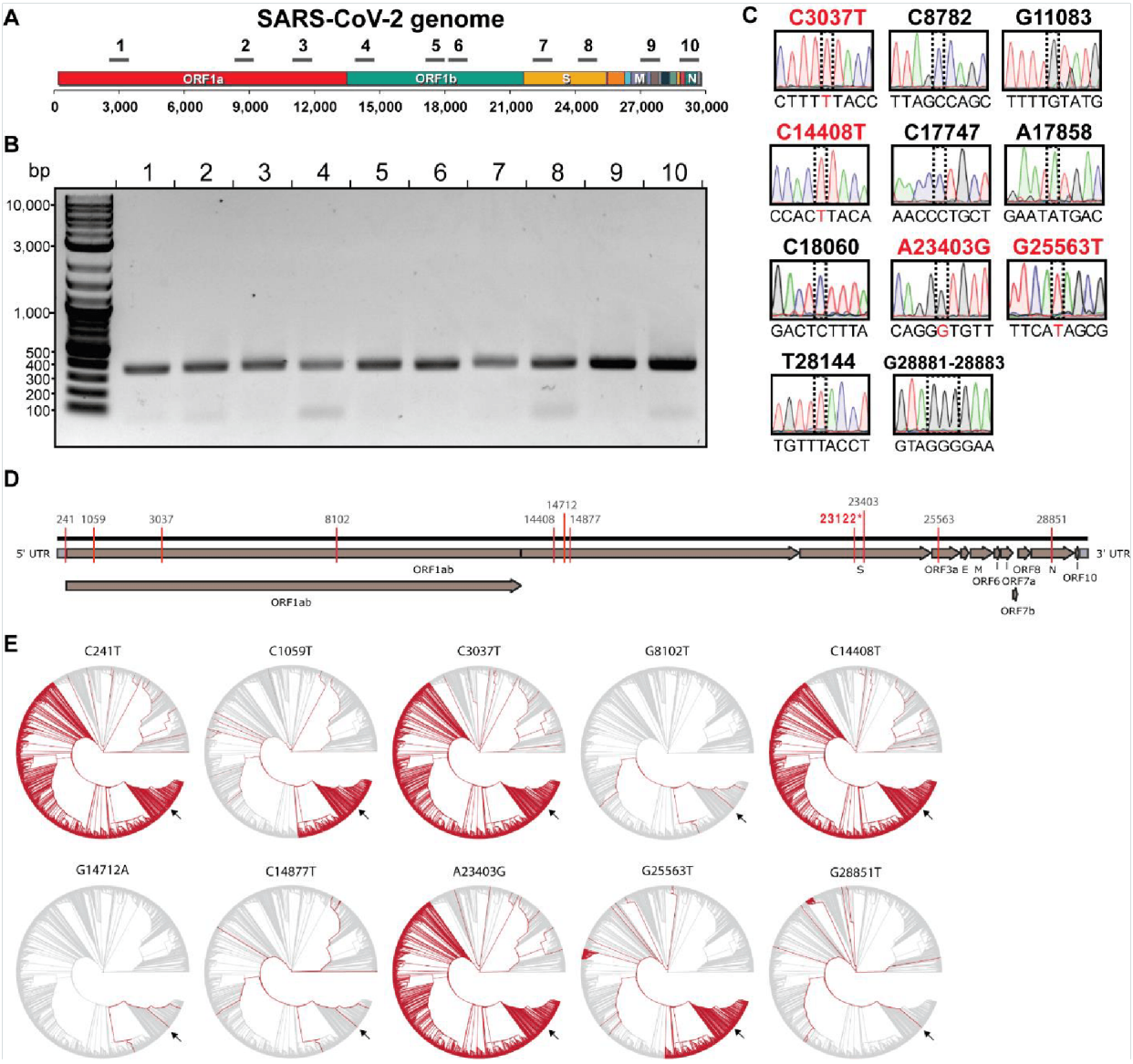
50 Rambaut, A. & Drummond, A. (2012).

51 Pagès H, A. P., Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. *R package version 2.52.0* (2019).

Supplemental Figures



Supplemental Figure S6.1, related to Figure 6.1. Standard curves for absolute quantification of SARS-CoV-2 titers. Standard curves were generated for each RT-qPCR run. A 10-fold dilution series of positive control plasmid (IDT#10006625) encoding for SARS-CoV-2 N gene were used, three technical replicates at each dilution (green circles). Data was plotted as Cycle Threshold (Ct) on y-axis versus log10(copy number) on x-axis. Trend lines were fitted to the data by linear regression analysis in RStudio v1.2.1335, Linear equations and R2 values are shown for each standard curve. Red dots show mean Ct values for wastewater samples. Error bars represent standard deviation of Ct values.



Supplemental Figure S6.2, related to Figure 6.2. Sequencing and phylogenetics of SARS-CoV-2 in wastewater. (A) Map of the SARS-CoV-2 genome. The solid black lines indicate the approximate location of 10 amplicons spanning phylogenetically informative regions of the genome. (B) Agarose gel of the corresponding PCR products. (C) Ten PCR products shown in B were gel-purified and used to sequence 13 polymorphic sites in SARS-CoV-2 genome [44]. Sanger traces (9-bp windows) are shown for each polymorphic site (dotted boxes). Nucleotide numbering is based on the reference genome Wuhan-Hu-1/2019 [25]. Highlighted in red are nucleotide substitutions relative to the reference. (D) Genetic map of SARS-CoV-2 with mutations present in Bozeman wastewater sequence highlighted. *Variant A23122T was further filtered out as an artifact of sequencing library preparation (see explanation in the main text). (E) Phylogenetic tree with each mutation present in Bozeman wastewater mapped in red. The SARS-CoV-2 sequence from Bozeman wastewater is indicated by black arrow.

Supplemental Tables

Sampling date	SARS-CoV-2 RNA concentration estimated with N1, copies / L*	SARS-CoV-2 RNA concentration estimated with N2, copies / L*	Total daily flow volume, L
March 30, 2020	528.9 ± 249	665.6 ± 261.5	22,871,458
April 1, 2020	432.7 ± 66.3	1710.8 ± 179.2	24,097,932
April 3, 2020	254.7 ± 5.4	836.3 ± 144.6	23,287,853
April 6, 2020	57.6 ± 38	257.4 ± 93.8	23,882,163
April 8, 2020	46 ± 9.2	23.5 ± 14.6	23,628,540
April 15, 2020	76 ± 24	46.7 ± 4.2	22,761,681
April 22, 2020	44.6 ± 5.6	22.2 **	22,492,917
April 29, 2020	43.5 ± 34.4	24.6 ± 1.5	22,273,363
May 6, 2020	0 ***	0	22,720,042
May 13, 2020	0	0	22,473,990
May 20, 2020	0	0	22,750,325
May 27, 2020	0	0	22,807,106
June 3, 2020	5400.5 ± 1046.1	6116.2 ± 1034.8	22,121,947
June 5, 2020	4174.6 ± 749.1	5624.6 ± 2509.2	21,754,762
June 8, 2020	2464.2 ± 502.4	1429 ± 322.4	21,016,606
June 10, 2020	971.9 ± 21.5	1947.9 ± 62.4	23,102,368
June 12, 2020	1762.7 ± 733.1	2291.4 ± 680.9	22,591,338

*Data shown as mean ± sem, n = 3
** Only one of three replicates showed amplification
*** Samples tested negative and thus were called 0 copies/L

Supplemental Table S6.1, related to Fig 6.1. Total wastewater flow volumes and SARS-CoV-2 RNA concentration in composite samples.

PCR product	Primer	Sequence (5'-3')
1	nCoV-2019_10_LEFT	TGAGAAGTGCTCTGCCTATACAGT
	nCoV-2019_10_RIGHT	TCATCTAACCAATCTTCTTCTTGCTCT
2	nCoV-2019_29_LEFT	ACTTGTGTTCCTTTTTGTTGCTGC
	nCoV-2019_29_RIGHT	AGTGTACTCTATAAGTTTTGATGGTGTGT
3	nCoV-2019_37_LEFT	ACACACCACTGGTTGTTACTCAC
	nCoV-2019_37_RIGHT	GTCCACACTCTCCTAGCACCAT
4	nCoV-2019_48_LEFT	TGTTGACACTGACTTAACAAAGCCT
	nCoV-2019_48_RIGHT	TAGATTACCAGAAGCAGCGTGC
5	nCoV-2019_59_LEFT	TCACGCATGATGTTTCATCTGCA
	nCoV-2019_59_RIGHT	AAGAGTCCTGTTACATTTTCAGCTTG
6	nCoV-2019_60_LEFT	TGATAGAGACCTTTATGACAAGTTGCA
	nCoV-2019_60_RIGHT	GGTACCAACAGCTTCTCTAGTAGC
7	nCoV-2019_77_LEFT	CCAGCAACTGTTTGTGGACCTA
	nCoV-2019_77_RIGHT	CAGCCCCTATTAAACAGCCTGC
8	nCoV-2019_84_LEFT	TGCTGTAGTTGTCTCAAGGGCT
	nCoV-2019_84_RIGHT	AGGTGTGAGTAAACTGTTACAAACAAC
9	nCoV-2019_93_LEFT	TGAGGCTGGTTCTAAATCACCCA
	nCoV-2019_93_RIGHT	AGGTCTTCCTTGCCATGTTGAG
10	nCoV-2019_95_LEFT	TGAGGGAGCCTTGAATACACCA
	nCoV-2019_95_RIGHT	CAGTACGTTTTTGCCGAGGCTT

Supplemental Table S6.2, related to Methods. Primer design used for RT-PCR (Supplemental Figure S6.2A and B)

Materials and Methods

Wastewater Samples

Wastewater samples were collected at the Bozeman Water Reclamation Facility (BWRF) that receives and treats domestic, commercial, and industrial wastewater from the City of Bozeman, Montana (USA). Wastewater is sourced from the city limits (~60 km^2; 49,831 population) with an average flow rate of ~2.31 3 10^4 m^3 / d. Composite samples were collected from raw influent with automatic flow proportional sampler Liquistation CSF34 (Endress+Hauser) located at the entrance to the facility downstream of a rock trap. Autosampler was set to collect 150 mL of influent per 150'000 gal of flow (~5.68 x 10^5 L) 7 AM to 7 AM. During collection temperature was kept +2º to +6ºC, and samples were stored at +4ºC before processing (2-3 h). The composite sample was subsampled in three 500 mL aliquots. No permissions were required for collection of the wastewater.

Symptom Onset Data and Clinical Test Results

Suspected cases of COVID-19 were tested in a CLIA lab and instructed to self-quarantine until notified of the RT-qPCR test results. All laboratory confirmed positive cases of COVID-19 were contacted via telephone by local public health nurses to complete contact tracing. During this interview, the nurses collected recorded symptoms, symptom onset date, travel history, contact with other known laboratory confirmed cases, close contacts and activities on the two days before symptom onset up until notification of a positive test. Data collection was conducted as part of a public health response. Information on COVID-19 patients COVID-19 (gender, age, disease severity, etc) is not available. The study was reviewed by the Montana State University Institutional Review Board (IRB) For the Protection of Human Subjects (FWA 00000165) and

was exempt from IRB oversight in accordance with Code of Federal regulations, Part 46, section 101. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Wastewater Sample Processing and RNA Extraction

Each wastewater sample (500 mL) was sequentially filtered through 20 mM, 5 mM (Sartorius Biolab Products) and 0.45 mM (Pall Corporation) membrane filters and concentrated down to 150-200 mL using Corning Spin-X UF concentrators with 100 kDa molecular weight cut-off. Total RNA from concentrated samples was extracted with RNeasy Mini Kit (QIAGEN) and eluted with 40 mL of RNase free buffer. This RNA was used as a template for RT-qPCR.

Reverse Transcription Quantitative PCR (RT-qPCR)

RT-qPCR was performed using two primers pairs (N1 and N2) and probes from 2019-nCoV CDC EUA Kit (IDT#10006606). SARS-CoV-2 in wastewater was detected and quantified using one-step RT-qPCR in ABI 7500 Fast Real-Time PCR System according to CDC guidelines and protocols (https://www.fda.gov/media/134922/download). 20 mL reactions included 8.5 mL of Nucleasefree Water, 1.5 mL of Primer and Probe mix, 5 mL of TaqPath 1-Step RT-qPCR Master Mix (ThermoFisher, A15299) and 5 mL of the template. Nuclease-free water was used as negative template control (NTC).

Amplification was performed as follows: 25ºC for 2 min, 50ºC for 15 min, 95ºC for 2 min followed by 45 cycles of 95ºC for 3 s and 55ºC for 30 s. To quantify viral genome copy numbers in the samples, standard curves for N1 and N2 were generated using a dilution series of a positive template control (PTC) plasmid (IDT#10006625) with concentrations ranging from 10 to 10,000 copies per reaction. Three technical replicates were performed at each dilution. The

limit of detection was 10 copies of the control plasmid. The NTC showed no amplification over the 40 cycles of qPCR.

Run data was analyzed in SDS software v1.4 (Applied Biosystems). Threshold cycle (Ct) values were determined by manually adjusting the threshold to fall within exponential phase of the fluorescence curves and above any background signal. Ct values of PTC dilutions were plotted against $\log_{10}$(copy number) to generate standard curves. Linear regression analysis was performed in RStudio v1.2.1335 and the trend line equation (Ct = [*slope*] 3 [$\log_{10}$(copy number)] + b) was used to calculate copy numbers from mean Ct values of technical replicates for each biological replicate. Primer efficiencies calculated as $E = \left(10^{\left(\frac{-1}{[slope]}\right)} - 1\right) x\ 100\%$ were 150.36 ± 12.13% for N1 and 129.45 ± 25.5% for N2 (n = 7 runs, mean ± sd).

RT-PCR and SARS-CoV-2 Genome Sequencing

Reverse transcription was performed with 10 mL of RNA from SARS-CoV-2 positive wastewater sample using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) with 10X RT Random Primers (Applied Biosystems) according to the supplier's protocol. Approximately 2000 viral RNA copies were used as an input for reverse transcription (estimated with qPCR).

The amplicon library for SARS-CoV-2 whole genome sequencing on Oxford Nanopore was generated as described in protocol developed by ARTIC Network (https://artic.network/ncov-2019).[22,26] Briefly, V3 primer pools containing 110 and 108 primers were used for the multiplex PCR (https://artic.network/ncov-2019). PCR reactions were performed using Q5 High-Fidelity DNA Polymerase (New England Biolabs) with the following thermocycling conditions: 98ºC for 2min, 35 cycles of 98ºC for 15 s and 65ºC for 5 min, 35

cycles. Two resulting amplicon pools were combined and used for library preparation pipeline that included end preparation and Nanopore adaptors ligation. 20 ng of final library DNA was loaded onto the MinION flowcell for sequencing. A total of 304.77 Mb of raw sequencing data was collected.

PCR products used for Sanger sequencing were generated with a subset of primers from ARTIC V3 pools (Table S6.2). PCR reactions were performed as described above. PCR products were analyzed on 1% agarose gels stained with SYBR Safe (Thermo Fisher Scientific), remaining DNA was purified using DNA Clean & Concentrator kit (Zymo Research) and sent to Psomagen for Sanger sequencing. Each PCR product was sequenced with both forward and reverse primers used for PCR.

SARS-CoV-2 Genome Assembly

Nanopore raw reads (304.77 Mb) were base-called with MinKNOW software in high-accuracy mode. Successfully base-called reads (273.8 Mb) were further analyzed using the ARTIC bioinformatic pipeline for COVID-19 (https://artic.network/ncov-2019). Consensus sequence was generated with minimap2 and single nucleotide variants were called with nanopolish (both integrated in the pipeline) relative to Wuhan-Hu-1/2019 reference genome (GenBank: MN908947.3).[25,43,44] The resulting assembly had nearly complete genome coverage (98.51%) with 6,875X average sequencing depth. Regions of the genome that were not captured by this sequencing method include 5' and 3' ends of the genome and a stretch of 170 nucleotides (22,346 – 22,515 nucleotide positions in reference genome), presumably due to amplicon drop-out. Consensus genome sequence was deposited to GISAID: EPI_ISL_437434 (https://www.gisaid.org/).

Phylogenetic analysis

Phylogenetic analysis was performed by aligning the consensus sequence to 14,970 SARS-CoV-2 genomes retrieved from GISAID on 5/5/2020, 8:25:22 AM (https://www.gisaid.org/), using the FFT-NS-2 setting in MAFFT v7.429.[45,46] Columns composed of more than 70% gaps were removed with trimAl v1.2rev59.[47] A maximum-likelihood phylogenetic tree was constructed from this alignment using IQTree in the Augur utility of Nextstrain.[23] The APE v5.3 package in R was used to re-root the tree relative to RaTG13 bat coronavirus genome sequence,[48] and the tree was plotted using ggtree v3.10 package in R.[49] The subtree, visualized in Figure 6.2B, was rendered in FigTree v1.4.4.[50]

Position-specific Mutation Analysis

Position specific mutation analysis was conducted in R using the BioStrings package,[51] and chromatograms of Sanger sequencing reads were rendered in SnapGene (GSL Biotech; available at snapgene.com).

Quantification and Statistical Analysis

All statistical analyses were performed in RStudio v1.2.1335. Data in figures are shown as mean of three biological replicates (each with two technical replicates) ± standard deviation (sd). Estimated copy numbers in RT-qPCR reactions were used to calculate titers per liter of wastewater for each biological replicate. Viral RNA concentrations in the composite samples were normalized ($[\text{SARS-CoV-2 concentration}]_{\text{Normalized}} = [\text{SARS-CoV-2 concentration}] \times (\text{Daily flow} / \text{Average flow})$).

Correlation analysis was performed in RStudio using stats, astsa and spatialEco R packages. Symptom onset, positive COVID-19 tests and SARS-CoV-2 RNA concentration in

wastewater measured with N1 or N2 primers were fit to a local polynomial regression (LOESS method) using the poly.regression wrapper function from spatialEco package. Resulting models were used to impute missing values (day w/o reported cases, periods between wastewater sampling days). Interpolated data were used for correlation analysis that was separately performed for the surge and resurgence of SARS-Co-V-2. Surge boundaries were determined as earliest reported symptom onset (left boundary) and date with last reported positive test (right boundary). The interval (mid-April – mid-May), when SARS-CoV-2 RNA was not detectable and no reported cases/symptoms were reported, was not included in the analysis. LOESS curves for wastewater data were displaced relative to epidemiological curves in one day increments from 14 to +14 days and Pearson correlation coefficient was calculated for each shift using corr function from stats R package.

CHAPTER SEVEN

SARS-COV-2 GENOMIC SURVEILLANCE IDENTIFIES NATURALLY OCCURING TRUNCATION OF ORF7A THAT LIMITS IMMUNE SUPPRESSION

Contribution of Authors and Co-Authors

Manuscript in Chapter 7

Author: Artem Nemudryi

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript, wrote the manuscript

Co-Author: Anna Nemudraia

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Tanner Wiegand

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Joseph Nichols

Contributions: performed the research

Co-Author: Deann T. Snyder

Contributions: designed the research, performed the research

Co-Author: Jodi F. Hedges

Contributions: designed the research, performed the research

Co-Author: Calvin Cicha

Contributions: performed the research

Co-Author: Helen Lee

Contributions: performed the research

Co-Author: Karl K. Vanderwood

Contributions: performed the research

Co-Author: Diane Bimczok

Contributions: acquired samples

Co-Author: Mark A. Jutila

Contributions: wrote the manuscript

Co-Author: Blake Wiedenheft

Contributions: acquired samples, designed the research, wrote the manuscript

Manuscript Information

Artem Nemudryi, Anna Nemudraia, Tanner Wiegand, Joseph Nichols, Deann T. Snyder, Jodi F. Hedges, Calvin Cicha, Helen Lee, Karl K. Vanderwood, Diane Bimczok, Mark A. Jutila, and Blake Wiedenheft

Cell Reports

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Cell Press
Volume 35
Issue 9
1 June 2021
DOI: 10.1016/j.celrep.2021.109197

SARS-COV-2 GENOMIC SURVEILLANCE IDENTIFIES NATURALLY OCCURING TRUNCATION OF ORF7A THAT LIMITS IMMUNE SUPPRESSION

Artem Nemudryi,[1] Anna Nemudraia,[1] Tanner Wiegand,[1,3] Joseph Nichols,[1,3] Deann T. Snyder,[1] Jodi F. Hedges,[1] Calvin Cicha,[1] Helen Lee,[1] Karl K. Vanderwood,[2] Diane Bimczok,[1] Mark A. Jutila,[1] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Gallatin City-County Health Department, Bozeman, MT 59715, USA. [3]These authors contributed equally.

Abstract

Over 950,000 whole-genome sequences of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have been determined for viruses isolated from around the world. These sequences are critical for understanding the spread and evolution of SARS-CoV-2. Using global phylogenomics, we show that mutations frequently occur in the C-terminal end of ORF7a. We isolate one of these mutant viruses from a patient sample and use viral challenge experiments to link this isolate (ORF7a$^{\Delta 115}$) to a growth defect. ORF7a is implicated in immune modulation, and we show that the C-terminal truncation negates anti-immune activities of the protein, which results in elevated type I interferon response to the viral infection. Collectively, this work indicates that ORF7a mutations occur frequently, and that these changes affect viral mechanisms responsible for suppressing the immune response.

Introduction

The spillover of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) into the human population has resulted in a global pandemic of coronavirus disease 2019 (COVID-19) with more than 2.8 million deaths worldwide (https://coronavirus.jhu.edu/) (Andersen et al., 2020). Although much of the research on this virus is focused on the Spike protein, recent reports demonstrate that accessory proteins of SARS-CoV-2 might be involved in COVID-19 pathogenesis by modulating antiviral host responses (Young et al., 2020; Zhang et al., 2020). SARS-CoV-2 uses multiple strategies to evade host immunity. Accessory proteins ORF3b, ORF6, and ORF7a antagonize various steps of type I interferon (IFN-I) production and signaling, while the proposed function of ORF8 is downregulating antigen presentation (Konno et al., 2020; Lei et al., 2020; Miorin et al., 2020; Tan et al., 2020; Xia et al., 2020). To occlude signal transmission from IFN receptors, ORF7a subverts phosphorylation of STAT2, suppressing transcriptional activation of antiviral IFN-stimulated genes (ISGs) that can otherwise restrict viral replication (Martin-Sancho et al., 2020; Xia et al., 2020). Although the proposed function for ORF7a is intracellular, antibodies against ORF7a are elevated in the serum of COVID-19 patients (Hachim et al., 2020). Work that is currently under review indicates that recombinant ORF7a protein interacts with $CD14^{+}$ monocytes and triggers expression of pro-inflammatory cytokines, including interleukin-6 (IL-6) and tumor necrosis factor alpha (TNF-a).[1] However, it is unclear if these interactions happen in COVID-19 patients. Together, these data suggest that ORF7a plays a dual role in SARS-CoV-2 infection by modulating both the IFN and inflammatory responses.

Results

SARS-CoV-2 Genomic Surveillance Identifies Truncations of ORF7a

Genome sequencing has been used to track the rise and spread of new SARS-CoV-2 lineages over the course of the pandemic (https://www.gisaid.org/).[2,3] As part of this effort, we sequenced SARS-CoV-2 genomes isolated from patients in Bozeman, Montana (Figure 7.1A). In total, we determined 55 whole-genome sequences of SARS-CoV-2 viruses isolated from patients between April and July 2020 using an amplicon-based Nanopore protocol from the ARTIC network (https://artic.network/).[4,5] To place the local outbreak in the context of the global SARS-CoV-2 evolutionary trends, we aligned 55 whole-genome sequences isolated from patients in Bozeman to 4,235 genomes subsampled (10 genomes per month per country) from 181,003 SARS-CoV-2 genomes downloaded from the GISAID. The subsampling protocol removes redundancy and bias introduced by uneven distribution of the global SARSCoV-2 sequencing effort. The resulting alignment was used to build a phylogenetic tree (Figure 7.1B). Of the 55 SARS-CoV-2 genomes determined from patients in Bozeman, only 1 was derived from the WA1 lineage (lineage A in Figure 7.1B) that was introduced to Washington state from Wuhan, China, in January 2020.[2,6] The remaining 54 genomes associate with clade B.1 and its offshoots, which are characterized by the D614G mutation in the spike and have prevailed globally since the spring of 2020 (Figure 7.1B).[7] The genetic variability in SARS-CoV-2 circulating in Bozeman (April to July 2020) represents 14 independent viral lineages, reflecting multiple introductions of the virus to the community from many different sources (Figure S7.1A).

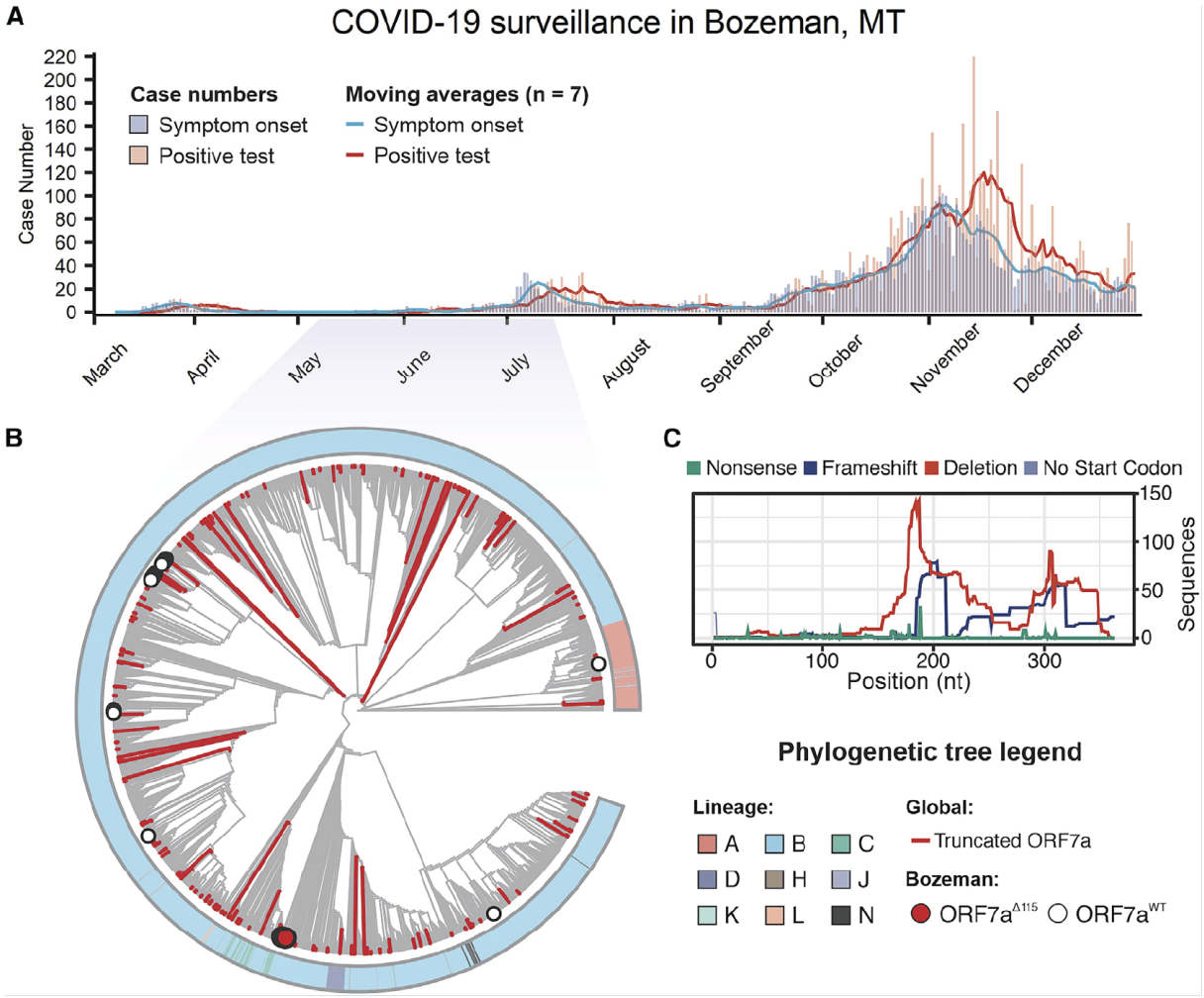


Figure 7.1. SARS-CoV-2 genomic surveillance identifies global reoccurrence of ORF7a truncations. (A) Symptom onset (purple) and PCR-based SARS-CoV-2 test results (coral) for patients in Bozeman, Montana, are shown with vertical bars. Seven-day moving averages, shown with lines, were used to indicate epidemiological trends. (B) Phylogenetic analysis of SARS-CoV-2 genomes sampled in Bozeman and globally. The tree was constructed from an alignment of 55 Bozeman samples and 4,871 genomes subsampled from GISAID. Subsampling was performed using Augur utility (https://nextstrain.org) by selecting 10 genomes per country per month since the start of the pandemic. Outer ring shows SARS-CoV-2 lineages assigned to genome sequences (Rambaut et al., 2020a). Major lineages include A (pink) that is associated with initial outbreak in China and B (blue) that emerged later in Europe. Minor lineages (i.e., C-N) are offshoots of lineage B. Red branches identify truncated ORF7a variants (n = 205) detected in the global data and merged into the alignment. The red dot highlights 7 of the 55 ORF7a variants that were isolated in Bozeman between April and July (2020). White dots highlight 48 viral genomes isolated in Bozeman that have wild-type ORF7a sequences. C) Distribution of different mutations that occur along the ORF7a coding sequence.

During the annotation of these genomes, we noticed a reoccurring (7 out of 55 genomes) 115-nt deletion in the gene-encoding accessory protein ORF7a (27,549–27,644 nt). The ORF7a$^{\Delta 115}$ mutation was found in patient swabs collected over a period of ~1.5 months. RT-PCR and Sanger sequencing verified that all these seven ORF7a$^{\Delta 115}$ variants are bona fide mutations and not a sequencing artifact (Figures S7.1B and S7.1C). In addition to the ORF7a$^{\Delta 115}$ deletion, these genomes also share 10 single-nucleotide variants (SNVs) compared with the Wuhan-Hu-1 reference genome. Seven of the 10 SNVs are found frequently worldwide and are a signature of the B1.1 lineage of SARS-CoV-2.[7] The remaining three mutations are rare, do not co-occur in any other genomes on GISAID, and lead to amino acid changes in ORF3a (Q38P, L95F) and N (R195I) proteins. Interestingly, one of the genomes in the ORF7a$^{\Delta 115}$ cluster has two additional SNVs that likely were acquired during circulation in the community. This virus was sampled from a patient 41 days after sampling the first ORF7a$^{\Delta 115}$ variant, which agrees well with estimated SARS-CoV-2 mutation rates (2 nt/month).[8]

To determine if the ORF7a$^{\Delta 115}$ genotype is unique to Bozeman, we downloaded an alignment of 180,971 SARS-CoV-2 genomes from GISAID, extracted ORF7a sequences, and determined their mutational profiles. In total, we identified 845 unique ORF7a gene variants that are different from the WuhanHu-1 reference sequence. Next, we looked for mutations with major effect on the ORF7a amino acid sequence. We identified 189 unique ORF7a variants (825 total sequences) with frameshifts, deletions, or missense mutations causing premature stop codons (Figure 7.1C). To understand the evolutionary history of these mutations, we integrated these genome sequences into our phylogenetic analysis. Detected ORF7a variants appear to have emerged independently on every continent and were not confined to any single lineage (Figure

7.1B). Next, we aligned translated ORF7a sequences to look for patterns. Most of the identified ORF7a mutations (126 out of 189) truncate the C terminus but preserve the N-terminal half of the protein (Figure 7.1C; Figure S7.1B). A portion (36.5%) of these truncated ORF7a variants appeared in two or more patient samples (116 maximum [max]). Although some of these ORF7a mutants appear to have arisen on multiple independent occasions, others come from genomes that form monophyletic clades, suggesting that viruses with ORF7a truncations are capable of transmission within a host population (Figure S7.1B).

ORF7a Truncation Has a Loss of Function Effect

ORF7a is a type I transmembrane (TM) protein with an N-terminal immunoglobulin-like (Ig-like) ectodomain, stalk, TM domain, and short cytosolic tail (Figure 7.2A). The two β sheets of the ectodomain are held together by two disulfide bonds (i.e., Cys23-Cys58 and Cys35-Cys67) (Figure 7.2B). The cytosolic tail of ORF7a contains a di-lysine (KRKTE) endoplasmic reticulum (ER) retrieval signal (ERRS) that mediates protein trafficking to the ER-Golgi intermediate compartment (ERGIC).[9]

The Δ115 nt mutation in ORF7a introduces a premature stop codon that eliminates b5, b6, and b7, two of the cystines that form disulfides (Cys58 and Cys67), the TM, and the cytosolic tail (Figures 7.2A and 7.2B). This truncation is expected to destabilize the protein structure and significantly impact protein function and trafficking in the host cell. To determine how loss of TM and sorting signal affects protein localization, we cloned and expressed FLAG-tagged wild-type and truncated ORF7a proteins in HEK293T-hACE2 cells. The wild-type ORF7a accumulated in the perinuclear region of the cell, which is consistent with previously reported ERGIC localization (Figure 2D) (Martin-Sancho et al., 2020; Nelson et al., 2005). In

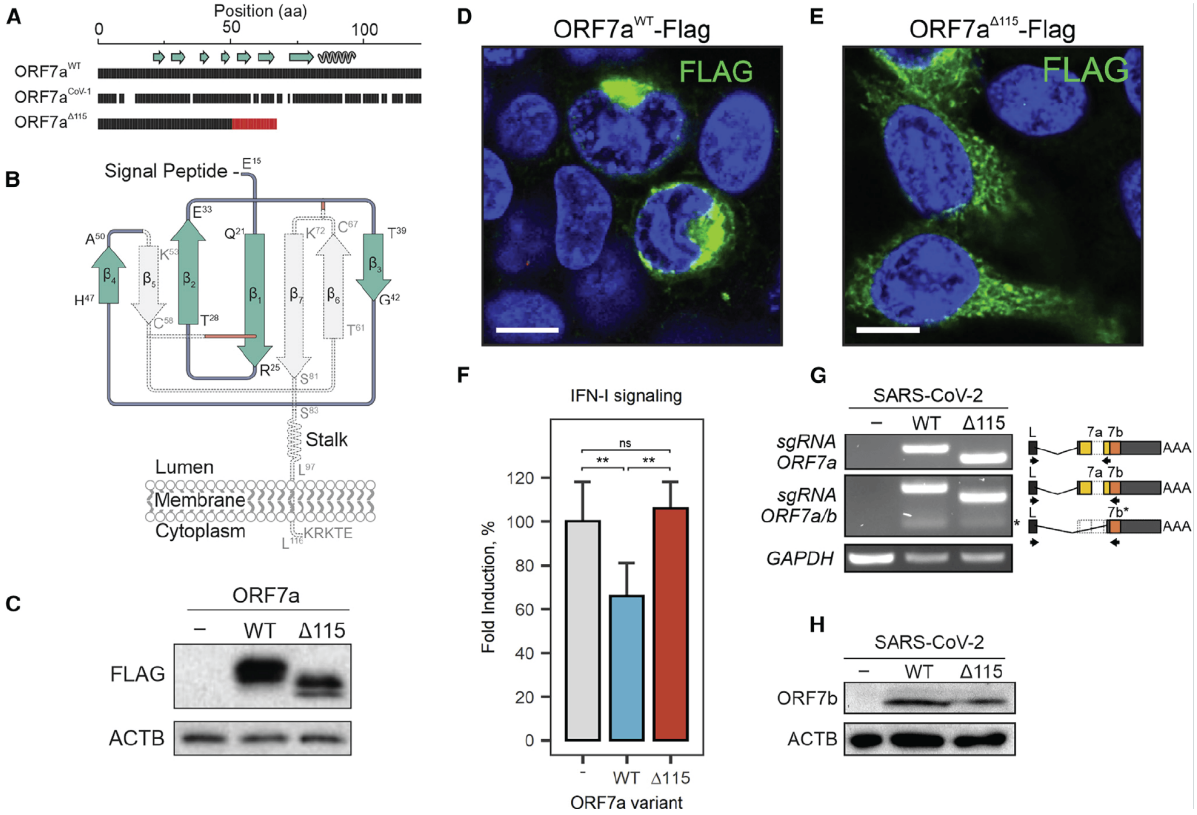


Figure 7.2. ORF7a truncation results in loss of function. (A) Amino acid (aa) sequence alignment of SARS-CoV-2 ORF7a^{WT}, ORF7a$^{\Delta115}$, and SARS-CoV-1 ORF7a. Gaps show non-matching positions; red shows 17-aa sequence resulting from a frameshift in the ORF7a mutant. Beta strands (arrows) and alpha helices (coil) are shown above the alignment.
(B) Diagram of SARS-CoV-2 ORF7a Ig-like fold. Disulfide bonds that stabilize the b sandwich structure are shown with red lines. The portion of the protein eliminated by the deletion is shown in gray. (C) C-terminally FLAG-tagged ORF7a^{WT} and ORF7a$^{\Delta115}$ were cloned and overexpressed in HEK293T-hACE2. Protein expression was confirmed with western blot using anti-FLAG antibody. b-Actin (ACTB) was used as a loading control. (D and E) FLAG-tagged ORF7a^{WT} (D) or ORF7a$^{\Delta115}$ (E) expressed in HEK293T-hACE2 cells. Immunostaining was performed using an anti-FLAG antibody (green). Cell nuclei were stained with Hoechst 33342 (blue). White scale bar is 10 mm. (F) HEK293T cells with integrated ISRE-luciferase reporter were transfected with pLV-mCherry (‘‘-’’), ORF7a^{WT}, or ORF7a$^{\Delta115}$plasmids. Transfected cells were treated with 5 ng/mL human recombinant IFN-a2b for 24 h, and induction of type I IFN signaling was measured with luciferase assay. Fold induction versus non- treated control was determined and normalized to mCherry control. Means (n = 6) were compared with one-way ANOVA ($p = 0.00162$). Pairwise comparisons were performed using post hoc Tukey’s test. Data are shown as mean ± SD. $^{**}p < 0.01$; $^{ns}p > 0.05$. (G) ORF7a and ORF7b sgRNAs were identified by RT-PCR. Lower molecular weight band corresponding to ORF7b sgRNA is indicated with asterisk (*). Diagram on the right shows primer (arrows) positions. Specificity of PCR products was confirmed with Sanger sequencing. *GAPDH* was used as a control for cDNA synthesis. (H) Lysates from SARS-CoV-2-infected cells were probed with antibodies raised against ORF7b protein. ACTB was used as a loading control.

contrast, the truncated ORF7a is distributed throughout the cytoplasm and does not associate with specific subcellular compartments, which is consistent with the loss of the TM domain and the ERRS signals required for protein targeting (Figure 7.2E).

Extent of the truncation and the defect in intracellular targeting suggests possible loss of ORF7a function. Along with several SARS-CoV-2 accessory proteins, ORF7a has been implicated in suppression of host IFN response to the infection.[10] To test the effect of the truncated ORF7a protein on IFN signaling, we generated reporter HEK293T cells with integrated luciferase reporter gene controlled by the interferon-stimulated response element (ISRE). In these cells, activation of interferon response drives reporter expression, which is quantified using a luciferase assay. To test immune suppression, we overexpressed the wild-type and the deletion mutant of ORF7a (ORF7a$^{\Delta 115}$) from plasmids and the stimulated cells with IFNa2b. The ORF7a WT protein suppressed IFN response by 34% ($p = 0.0074$) compared with cells transfected with control plasmid, which is consistent with recent results by Xia et al. (2020).[10] In contrast, the ORF7a$^{\Delta 115}$ protein failed to suppress IFN response and showed no significant difference to a control (mCherry) plasmid (Figure 7.2F). These results indicate that the Δ115 mutation in *ORF7a* has a loss-of-function effect that negates antagonism of IFN signaling.

ORF7a Mutation Has No Collateral Effect on ORF7b

The ORF7a gene overlaps with the downstream ORF7b gene (Figure S7.2). To determine if the Δ115 mutation in ORF7a impacts ORF7b, we first examined whether it eliminates a transcription-regulatory sequence (TRS) that is required for subgenomic RNA (sgRNA) synthesis. None of the TRSs identified via direct RNA sequencing (RNA-seq) of SARS-CoV-2 overlap with the Δ115 mutation site (Kim et al., 2020),[11] suggesting that ORF7b transcription is

not affected (Figure S7.2). To confirm this prediction, we used RT-PCR with primers that detect ORF7a and ORF7b sgRNAs (Figure 7.2G). In this assay, we infected 293T-hACE2 cells (MOI = 0.05) with ORF7a^{WT} or ORF7a$^{\Delta 115}$ viral strains and extracted total RNA from cells at 24 hours post-infection (hpi). Both viruses produced specific RT-PCR products corresponding to ORF7a and ORF7b sgRNAs. Finally, to verify that the ORF7a$^{\Delta 115}$ variant does not impact ORF7b translation, we probed cell lysates 24 hpi with an anti-ORF7b antibody (Figure 7.2H). We detected ORF7b protein in both viral strains with band intensities that suggest similar expression levels. Collectively, these data indicate that the ORF7a$^{\Delta 115}$ mutation has no collateral effect on the adjacent gene.

ORF7a$^{\Delta 115}$ Isolate Displays a Replication Defect

To investigate the phenotype of SARS-CoV-2 Δ115 variant, we infected 293T-hACE2 and Vero E6 cells with ORF7a^{WT} and ORF7a$^{\Delta 115}$ viruses at an MOI of 0.05 and measured viral RNA replication using qRT-PCR (Figures 7.3A and 7.3B). In this assay, we used viral strains sharing the same haplotype (C241T, C3037T, C14408T, A23403G) that defines the SARS-CoV-2 B.1 lineage and its derivatives (Figure S7.3A).[7] Over the course of 120 h, ORF7a^{WT} viral RNA level increased ~960-fold in supernatants from Vero E6 cells (Figure 7.3A) and ~270-fold in supernatants from 293T-hACE2 cells (Figure 7.3B). The ORF7a$^{\Delta 115}$ virus displayed limited replication that resulted in only 18.6-fold and 7-fold RNA level increase at 120 hpi in VeroE6 and 293T-hACE2, respectively. This growth defect reaches statistical significance starting at 6 hpi in Vero E6 cells ($p = 0.0135$) and 24 hpi in 293T-hACE2 cells ($p = 6.2 \times 10^{5}$).

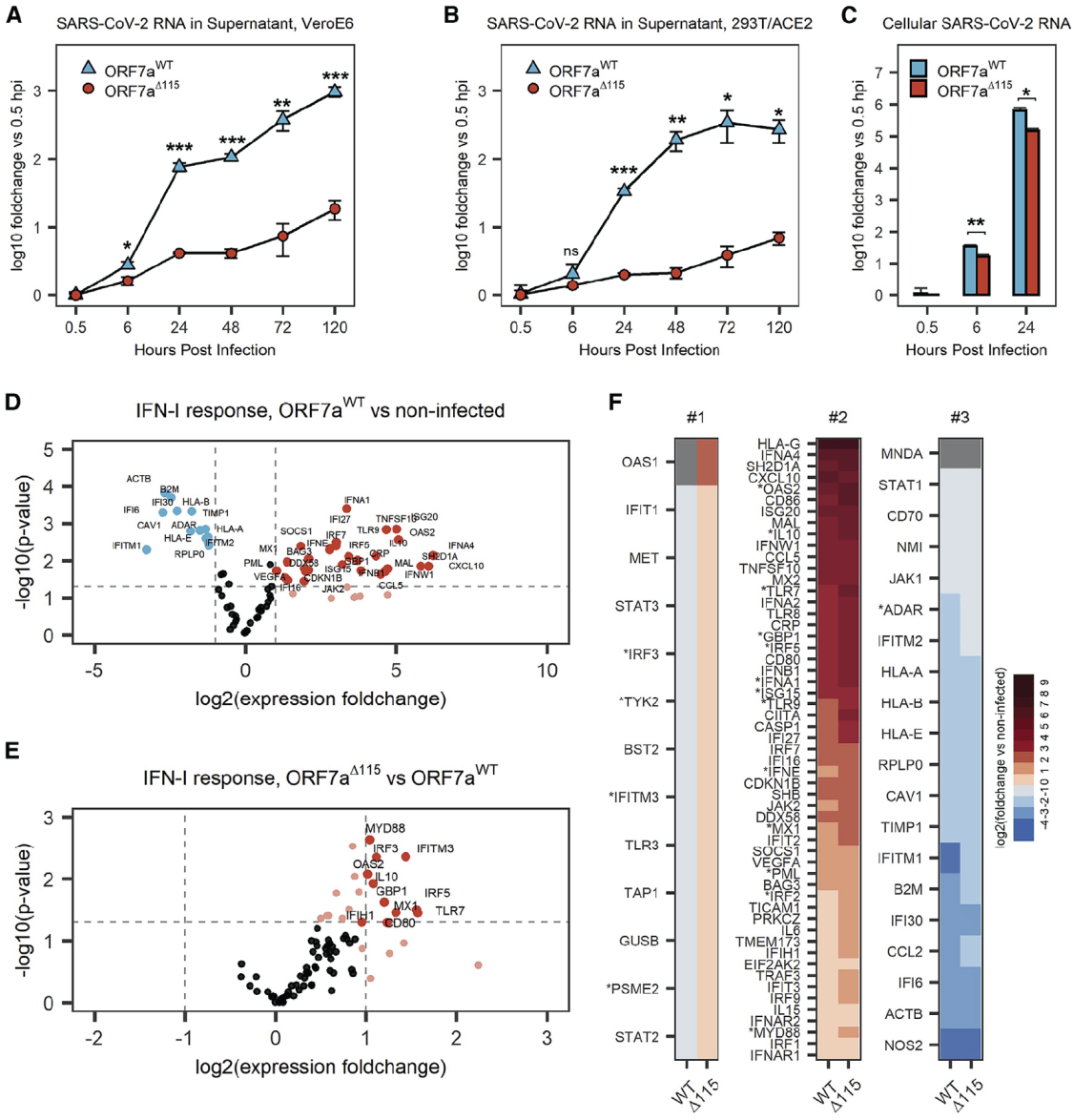


Figure 7.3. The ORF7a$^{\Delta 115}$ mutation results in replication and immune suppression defects (A and B) VeroE6 (A) or HEK293T-hACE2 (B) cells were infected with ORF7a^{WT} or ORF7a$^{\Delta 115}$SARS-CoV-2 strain at MOI = 0.05. Viral RNA was measured in supernatant using qRT-PCR at different time points after infection. Measured RNA levels were normalized (using DCt method) to RNA levels at 0.5 hpi. (C) Total RNA was extracted from infected HEK293T-hACE2 cells, and viral RNA was measured. The DDCt method was used to normalize viral RNA levels to host RNA (*ACTB*) and 0.5 hpi time point. Data in (A)–(C) are presented as mean ± SD of three biological replicates. Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001, or ^{ns}p > 0.05 (no significant difference). (D) Volcano plot showing IFN-I response in 293T-hACE2 cells infected with ORF7a^{WT} SARS-CoV-2 strain at MOI = 0.05. Expression of IFN-I response genes was studied 24 hpi using qRT-PCR array targeting 91 human transcripts (88 targets and 3 references). Experiment was performed in three biological replicates. Dashed lines show

regulation (≥ 2-fold) and statistical significance ($p < 0.05$) thresholds. Each dot represents mean ($n = 3$) normalized expression of a single gene relative to non-infected host. Genes that passed the threshold are labeled. (E) Volcano plot showing IFN-I response in ORF7a$^{\Delta 115}$ versus ORF7a^{WT} infection. (F) Data shown in (D) and Figure S7.3A were plotted as a heatmap. Genes were classified into three groups. Group #1 genes are oppositely regulated between the two viral strains. Group #2 genes are upregulated by both ORF7a^{WT} and ORF7a$^{\Delta 115}$. Group #3 genes are downregulated in both. Genes that have statistically significant difference in expression between two viral strains are marked with asterisk. Gray shows genes with no detectable expression over 40 cycles of qRT-PCR.

To determine if this phenotype results from a defect in viral egress only or if viral RNA replication in the host is also affected, we examined early steps in the infection. We extracted total RNA from the infected 293T-hACE2 cells at 0.5, 6, and 24 hpi and measured SARS-CoV-2 RNA with qRT-PCR. Compared with wild type, ORF7a$^{\Delta 115}$ RNA level was reduced 2.1-fold inside the host at 6 hpi ($p = 0.002$) and 4.3fold at 24 hpi ($p = 0.02$) (Figure7. 3C). This decrease in viral RNA in the host cell differs from the corresponding decrease in the supernatant RNA at 24 hpi (4.3-fold versus 16.8-fold, respectively; Figures 7.3B and 7.3C). This disparity suggests that observed growth defect in the SARS-CoV-2$^{\Delta 115}$ variant might result from defects in genomic RNA replication and transcription, as well as in viral egress.

ORF7a$^{\Delta 115}$ Virus Induces Elevated IFN-I response *in vitro*

Suppressed or delayed IFN responses characterize SARS-CoV2 infections.[12] The C-terminal truncation of ORF7a limits suppression of the type I IFN signaling (Figure 7.2F). Therefore, we hypothesized that ORF7a$^{\Delta 115}$ virus isolated from the community is deficient in antagonizing IFN response.

To test this hypothesis, we measured expression of type I IFN response genes upon infection with ORF7a^{WT} and ORF7a$^{\Delta 115}$ strains using a qRT-PCR array targeting 91 human transcripts. The array included three reference genes that we used to normalize expression of 88

target genes. In agreement with recent transcriptomic studies, we detected a type I IFN-mediated antiviral response after infecting HEK293T-hACE2 with SARS-CoV-2 isolates (Figure 7.3D; Figure S7.3B).[13,14] Out of 88 type I IFN response genes included in the qRT-PCR array, 42 were differentially expressed (≥ 2-fold change, $p < 0.05$) in ORF7a^{WT}-infected cells and 55 in ORF7a$^{\Delta 115}$ infection, as compared with the non-infected control (Figure 7.3D; Figure S7.3B). Both viruses stimulated expression of multiple type I interferon genes, antiviral ISGs, ILs, and proinflammatory cytokines.

To understand the effect of ORF7a$^{\Delta 115}$ on the host type I IFN response, we compared expression of genes targeted with qRT-PCR array between the two viruses. In general, the ORF7a$^{\Delta 115}$ variant induces elevated type I IFN response to the infection (Figures S7.3C and S7.3D). The median expression fold change of 88 genes targeted in the qRT-PCR array was 1.72 in ORF7a^{WT} and 2.51 in ORF7a$^{\Delta 115}$ virus ($p = 1.4E-12$; Figure S7.3C). Analysis of genes differentially expressed between the two viral infections results in an asymmetric volcano plot (Figure 7.3E). A subset of ISGs had expression levels that are significantly different ($p < 0.05$) between the two viral infections (Figures 7.3E and 7.3F). This subset includes sensors (*TLR7*), signal transducers (*MYD88, OAS2*), transcriptional regulators (*IRF3*, *IRF5*), and restriction factors (*GBP1*, *IFITM3*, *MX1*) known to combat infections by RNA viruses.[15] One of these differentially expressed ISGs (*IFITM3*) restricts SARS-CoV-2 entry.[16,17] Additionally, polymorphisms in *IFITM3*, *MX1*, and *TLR7* are linked to the severity of COVID-19.[18-20] Upregulation of ISGs indicates that the ORF7a$^{\Delta 115}$ virus is inept at antagonizing the IFN-I response to the infection, which is consistent with results from our ISRE reporter assay (Figure 7.2F).

Discussion

In this study, we use genomic surveillance and phylogenetics to identify ORF7a variants that have emerged globally throughout the SARS-CoV-2 pandemic. Local occurrence of ORF7a mutations has been reported by others.[21-23] However, the effect of these mutations on viral fitness and host immune responses has not been investigated. Here, we isolated a virus with a deletion mutation in ORF7a (ORF7a$^{\Delta 115}$) and show that this genotype results in an *in vitro* growth defect. This growth defect is associated with elevated IFN response to SARS-CoV-2.

Interferon systems are a frontline of host defense that signals infection and elicits an antiviral state.[24] Pretreatment with type I, type II, and type III IFNs restricts SARSCoV-2 infection and replication in cell culture.[25,26] SARS-CoV-2 deploys multiple proteins (e.g., nsp6, nsp13, ORF6, ORF7a, ORF7b, and ORF9b) to shut down IFN signaling and dampen innate immune responses,[10,12,27] in particular, when overexpressed ORF7a subverts STAT2 phosphorylation, blocking IFN-dependent transcriptional activation of antiviral ISGs.[10] Both type I and III IFNs signal through the JAK-STAT pathway and overlap substantially in the transcriptional responses they induce.[28] Therefore, ORF7a likely suppresses both IFN types; however, its role in suppressing the type III IFN response has yet to be verified.[10] Using the ISRE reporter system, we show that C-terminal truncation negates the ability of ORF7a to suppress IFN signaling (Figure 7.2). The ORF7a$^{\Delta 115}$ variant shows limited replication in HEK293T-hACE2, as well as in Vero E6, the latter of which does not express type I IFN genes.[29,30] Although type I IFN genes are deleted in Vero E6 cells, the type III IFNs are intact and are activated upon infection with RNA viruses, which results in ISG.[31,32] ISGs exert anti-viral activities that combat the infection.[15] We hypothesize that the replication defect of the

ORF7a$^{\Delta 115}$ strain in Vero E6 and 293T-hACE2 cells results from an IFN-dependent response that the virus fails to suppress.

Several groups have replaced ORF7a with a reporter gene.[33-35] Thi Nhu Thao et al. (2020) show that complete deletion of ORF7a reduces replication of the synthetic virus, which agrees well with growth defect in the ORF7a$^{\Delta 115}$ strain.[34] However, two other studies have not observed this effect. We believe that our results with a naturally occurring ORF7a deletion provide valuable insight to this discussion. Finally, it should be noted that different mutations (i.e., complete deletion versus partial) might have a different outcome for the viral phenotype.

Cytokine profiling and RNA-seq of clinical samples reveal suppressed or delayed IFN responses in COVID-19 patients.[36-38] Dysregulated immune responses coupled with exacerbated inflammatory cytokines drive pathogenesis of the disease and correlate with its severity.[39-41] Here, we detected enhanced IFN signaling in cells infected with a naturally occurring ORF7a$^{\Delta 115}$ variant of SARS-CoV-2.

Given the importance of the IFN response in COVID-19, we anticipate that ORF7a mutations might affect SARS-CoV-2 pathogenicity. Clinical comparisons will be required to examine this potential effect on COVID-19 features.

Deletions in ORF7a, ORF7b, and ORF8 have previously been reported.[21-23,42] Similar deletions emerged in Middle East Respiratory Syndrome (MERS) and SARS-CoV, which were linked to viral attenuati.[43,44] A 382-nt deletion in ORF8 is associated with milder SARS-CoV-2 infection in patients.[45] Although this mutation affects clinical features of COVID-19, it has no evident effect on viral replication or host transcriptional responses in human cell culture.[46]

Several studies illustrate how persistent SARS-CoV-2 infections in immunocompromised hosts associate with accelerated viral evolution.[47,48] In these hosts there is less selective pressure on viral anti-immune proteins, which might permit loss-of-function mutations in accessory genes. The ORF7a$^{\Delta 115}$ variant that we isolated was present in only seven genomes over a course of ~1.5 months in a local community and then disappeared. We hypothesize that deletions in accessory genes, such as ORF7a$^{\Delta 115}$, might emerge in immunocompromised patients but do not persist in the immunocompetent population. Similarly, a deletion mutation in ORF8 (i.e., D382), which was associated with milder COVID-19 disease, disappeared soon after it was discovered.[45] However, founder effect or mutations that co-occur with other mutations that increase transmissibility (e.g., spike mutations) can result in fixation of an otherwise attenuating mutation. A new more transmissible lineage of SARS-CoV-2 (i.e., B.1.1.7) has recently.[49,50] Aside from changes in the Spike protein, this lineage also includes a premature stop codon in the accessory ORF8 that truncates most of the protein.[51] This lineage might illustrate the second scenario for fixing mutations in accessory genes.

Limitations of Study

The viral replication defect and cellular immune response assay presented here were performed using HEK293T-hACE2 and Vero E6 cells that are not the natural targets of SARS-CoV-2. However, both cell lines are permissive for SARS-CoV-2 infection and are used extensively to study efficacy of antiviral compounds,[52] virion assembly,[53] SARS-CoV-2 entry,[54] replication,[34] and anti-immune activities of SARS-CoV-2 proteins.[10] Although we acknowledge that our conclusions are limited to Vero E6 and 293T-hACE2 cells, we anticipate that the results

presented here will provide important context for clinical comparisons, similar to those recently published for ORF8.[45]

Studies of naturally occurring SARS-CoV-2 variants have limitations, including unavailability of isogenic controls. Here, we compared two viral isolates, ORF7a^{WT} and ORF7a$^{\Delta 115}$, that in addition to Δ115 differ at several other nucleotide positions (Figure S7.3A). Although ISRE reporter assay unambiguously links the immunosuppression defect to the truncation of ORF7a, it is not possible to rule out a role for these mutations on the viral life cycle.

Acknowledgements

We are grateful to members of Bozeman Health that provided de-identified patient samples, specifically C. Nero, D. Smoot, W. Wallace, C. Faurot-Daniels, V. Lawrence, and M. Blauvelt. We are also grateful to M. Flenniken, K. Daughenbaugh, and other members of the COVID task force at MSU for assistance establishing the COVID testing center. We thank the Lefcort lab for generous use of their fluorescent confocal microscope and Dr. Pincus for providing the HEK 293T-hACE2 cells. Research in the Wiedenheft lab is supported by the National Institutes of Health (1R35GM134867), the Montana State University Agricultural Experimental Station, the MJ Murdock Charitable Trust, the Gianforte Foundation, and the MSU Office of the Vice President for Research. We thank the GISAID's EpiFlu Database and contributing laboratories. The phylogenetic analysis in this paper would not have been possible without their willingness to share data. The graphical abstract was created with BioRender. B.W. is the founder of SurGene LLC and VIRIS Detection Systems Inc. B.W., A. Nemudryi, and A. Nemudraia are inventors on patents related to CRISPR-Cas systems and applications thereof.

References

1 Zhou, Z. *et al.* Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14(+) monocytes. *iScience* **24**, 102187, doi:10.1016/j.isci.2021.102187 (2021).

2 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571-575, doi:10.1126/science.abc0523 (2020).

3 Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e819, doi:10.1016/j.cell.2020.06.043 (2020).

4 Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature protocols* **12**, 1261-1276, doi:10.1038/nprot.2017.066 (2017).

5 Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*, doi:10.1101/2020.09.04.283077 (2020).

6 Holshue, M. L. *et al.* First Case of 2019 Novel Coronavirus in the United States. *N Engl J Med* **382**, 929-936, doi:10.1056/NEJMoa2001191 (2020).

7 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).

8 van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* **83**, 104351-104351, doi:10.1016/j.meegid.2020.104351 (2020).

9 Nelson, C. A., Pekosz, A., Lee, C. A., Diamond, M. S. & Fremont, D. H. Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein. *Structure* **13**, 75-85, doi:10.1016/j.str.2004.10.010 (2005).

10 Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep* **33**, 108234, doi:10.1016/j.celrep.2020.108234 (2020).

11 Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921 e910, doi:10.1016/j.cell.2020.04.011 (2020).

12 Lei, X. *et al.* Activation and evasion of type I interferon responses by SARS-CoV-2. *Nat Commun* **11**, 3810, doi:10.1038/s41467-020-17665-9 (2020).

13 Banerjee, A. *et al.* Experimental and natural evidence of SARS-CoV-2-infection-induced activation of type I interferon responses. *iScience* **24**, 102477, doi:10.1016/j.isci.2021.102477 (2021).

14 Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036-1045 e1039, doi:10.1016/j.cell.2020.04.026 (2020).

15 Schoggins, J. W. Interferon-Stimulated Genes: What Do They All Do? *Annu Rev Virol* **6**, 567-584, doi:10.1146/annurev-virology-092818-015756 (2019).

16 Shi, G. *et al.* Opposing activities of IFITM proteins in SARS-CoV-2 infection. *EMBO J* **40**, e106501, doi:10.15252/embj.2020106501 (2021).

17 Zang, R. *et al.* Cholesterol 25-hydroxylase suppresses SARS-CoV-2 replication by blocking membrane fusion. *Proc Natl Acad Sci U S A* **117**, 32105-32113, doi:10.1073/pnas.2012197117 (2020).

18 Andolfo, I. *et al.* Common variants at 21q22.3 locus influence MX1 and TMPRSS2 gene expression and susceptibility to severe COVID-19. *iScience* **24**, 102322, doi:10.1016/j.isci.2021.102322 (2021).

19 Zhang, Y. *et al.* Interferon-Induced Transmembrane Protein 3 Genetic Variant rs12252-C Associated With Disease Severity in Coronavirus Disease 2019. *J Infect Dis* **222**, 34-37, doi:10.1093/infdis/jiaa224 (2020).

20 Pati, A., Padhi, S., Suvankar, S. & Panda, A. K. Minor Allele of Interferon-Induced Transmembrane Protein 3 Polymorphism (rs12252) Is Covered Against Severe Acute Respiratory Syndrome Coronavirus 2 Infection and Mortality: A Worldwide Epidemiological Investigation. *J Infect Dis* **223**, 175-178, doi:10.1093/infdis/jiaa630 (2021).

21 Addetia, A. *et al.* Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. *J Clin Virol* **129**, 104523, doi:10.1016/j.jcv.2020.104523 (2020).

22 Holland, L. A. *et al.* An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J Virol* **94**, doi:10.1128/JVI.00711-20 (2020).

23 Rosenthal, S. H. *et al.* Identification of eight SARS-CoV-2 ORF7a deletion variants in 2,726 clinical specimens. *bioRxiv*, 2020.2012.2010.418855, doi:10.1101/2020.12.10.418855 (2020).

24 McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in infectious disease. *Nat Rev Immunol* **15**, 87-103, doi:10.1038/nri3787 (2015).

25 Mantlo, E., Bukreyeva, N., Maruyama, J., Paessler, S. & Huang, C. Antiviral activities of type I interferons to SARS-CoV-2 infection. *Antiviral Res* **179**, 104811, doi:10.1016/j.antiviral.2020.104811 (2020).

26 Miorin, L. *et al.* SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proc Natl Acad Sci U S A* **117**, 28344-28354, doi:10.1073/pnas.2016650117 (2020).

27 Jiang, H. W. *et al.* SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. *Cell Mol Immunol* **17**, 998-1000, doi:10.1038/s41423-020-0514-8 (2020).

28 Kotenko, S. V. & Durbin, J. E. Contribution of type III interferons to antiviral immunity: location, location, location. *J Biol Chem* **292**, 7295-7303, doi:10.1074/jbc.R117.777102 (2017).

29 Emeny, J. M. & Morgan, M. J. Regulation of the interferon system: evidence that Vero cells have a genetic defect in interferon production. *J Gen Virol* **43**, 247-252, doi:10.1099/0022-1317-43-1-247 (1979).

30 Osada, N. *et al.* The genome landscape of the african green monkey kidney-derived vero cell line. *DNA Res* **21**, 673-683, doi:10.1093/dnares/dsu029 (2014).

31 Stoltz, M. & Klingstrom, J. Alpha/beta interferon (IFN-alpha/beta)-independent induction of IFN-lambda1 (interleukin-29) in response to Hantaan virus infection. *J Virol* **84**, 9140-9148, doi:10.1128/JVI.00717-10 (2010).

32 Wang, C. *et al.* The Coronavirus PEDV Evades Type III Interferon Response Through the miR-30c-5p/SOCS1 Axis. *Front Microbiol* **11**, 1180, doi:10.3389/fmicb.2020.01180 (2020).

33 Hou, Y. J. *et al.* SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. *Cell* **182**, 429-446 e414, doi:10.1016/j.cell.2020.05.042 (2020).

34 Thi Nhu Thao, T. *et al.* Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. *Nature* **582**, 561-565, doi:10.1038/s41586-020-2294-9 (2020).

35 Xie, X. *et al.* An Infectious cDNA Clone of SARS-CoV-2. *Cell Host Microbe* **27**, 841-848 e843, doi:10.1016/j.chom.2020.04.004 (2020).

36 Arunachalam, P. S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210-1220, doi:10.1126/science.abc6261 (2020).

37 Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718-724, doi:10.1126/science.abc6027 (2020).

38 Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463-469, doi:10.1038/s41586-020-2588-y (2020).

39 Zhang, B. *et al.* Clinical characteristics of 82 cases of death from COVID-19. *PLoS One* **15**, e0235458, doi:10.1371/journal.pone.0235458 (2020).

40 Tay, M. Z., Poh, C. M., Renia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* **20**, 363-374, doi:10.1038/s41577-020-0311-8 (2020).

41 Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054-1062, doi:10.1016/S0140-6736(20)30566-3 (2020).

42 Su, Y. C. F. *et al.* Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *mBio* **11**, doi:10.1128/mBio.01610-20 (2020).

43 Chinese, S. M. E. C. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666-1669, doi:10.1126/science.1092002 (2004).

44 Muth, D. *et al.* Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci Rep* **8**, 15177, doi:10.1038/s41598-018-33487-8 (2018).

45 Young, B. E. *et al.* Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* **396**, 603-611, doi:10.1016/S0140-6736(20)31757-8 (2020).

46 Gamage, A. M. *et al.* Infection of human Nasal Epithelial Cells with SARS-CoV-2 and a 382-nt deletion isolate lacking ORF8 reveals similar viral kinetics and host transcriptional profiles. *PLoS Pathog* **16**, e1009130, doi:10.1371/journal.ppat.1009130 (2020).

47 Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912 e1909, doi:10.1016/j.cell.2020.10.049 (2020).

48 Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med* **383**, 2291-2293, doi:10.1056/NEJMc2031364 (2020).

49 Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, doi:10.1126/science.abg3055 (2021).

50 Andrew Rambaut, N. L., Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz, on behalf of COVID-19 Genomics Consortium UK (CoG-UK). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*, 2020).

51 Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*, 2020.2012.2030.20249034, doi:10.1101/2020.12.30.20249034 (2021).

52 Riva, L. *et al.* Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113-119, doi:10.1038/s41586-020-2577-1 (2020).

53 Klein, S. *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nat Commun* **11**, 5885, doi:10.1038/s41467-020-19619-7 (2020).

54 Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280 e278, doi:10.1016/j.cell.2020.02.052 (2020).

55 Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-D503, doi:10.1093/nar/gkz1031 (2020).

56 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

57 Harcourt, J. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 from Patient with Coronavirus Disease, United States. *Emerg Infect Dis* **26**, 1266-1273, doi:10.3201/eid2606.200516 (2020).

58 Loveday, E. K. *et al.* Effect of Inactivation Methods on SARS-CoV-2 Virion Protein and Structure. *Viruses* **13**, doi:10.3390/v13040562 (2021).

59 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome biology* **20**, 8-8, doi:10.1186/s13059-018-1618-7 (2019).

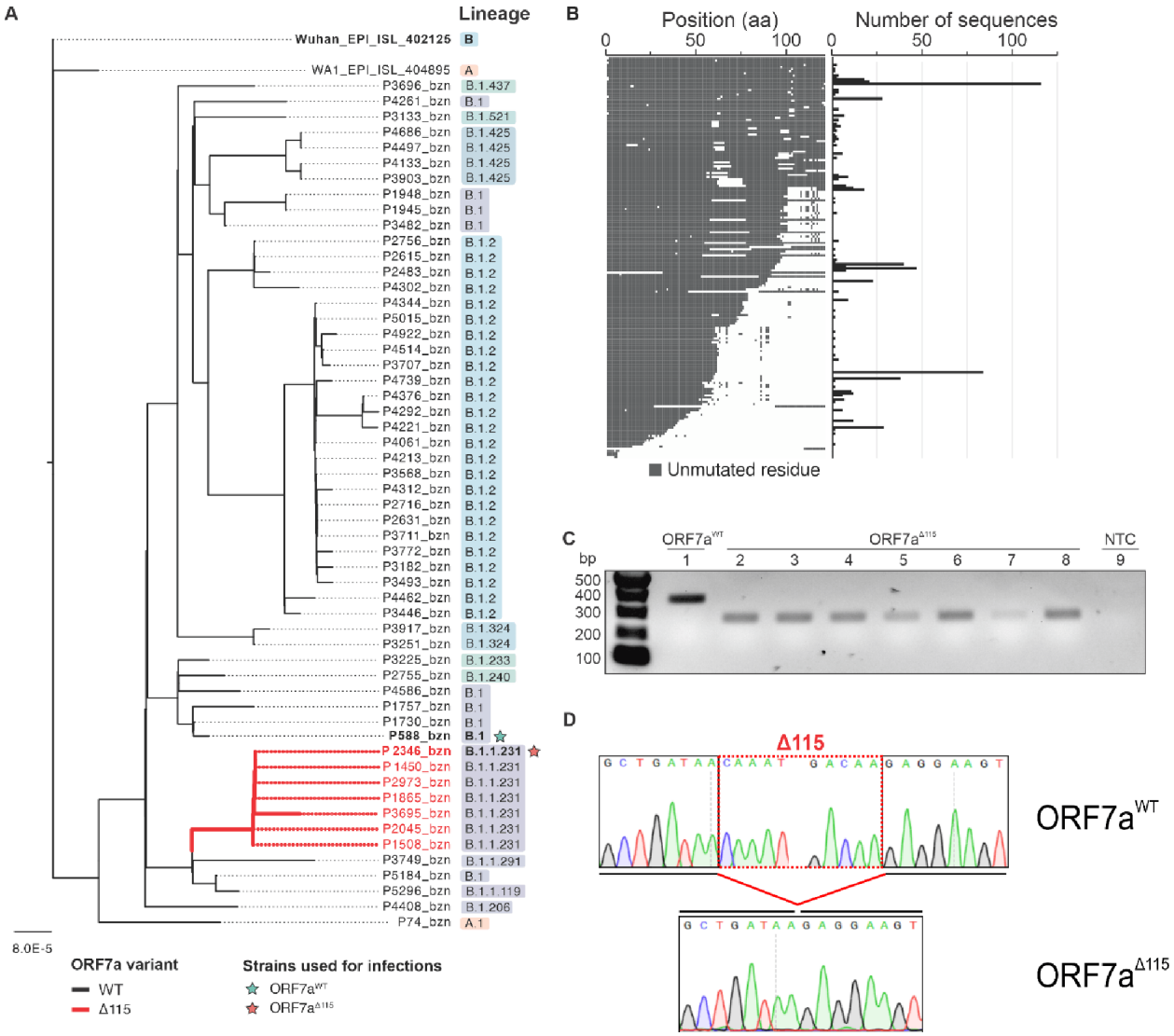
60 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

61 Quick, J. in *protocols.io* (ZappyLab, Inc., 2020).

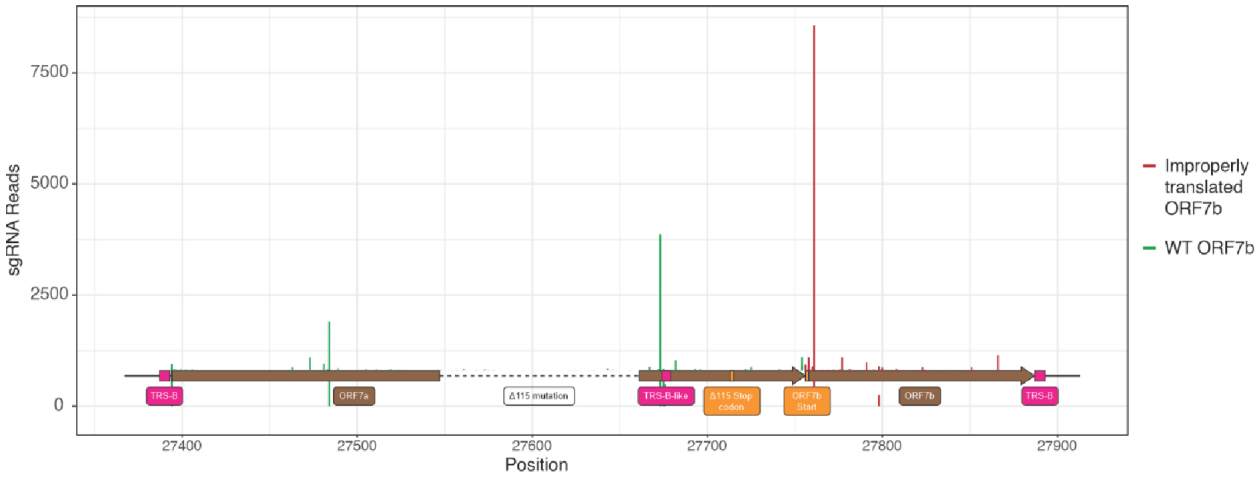
62 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

63 Lemieux, J. E. *et al.* Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, doi:10.1126/science.abe3261 (2021).

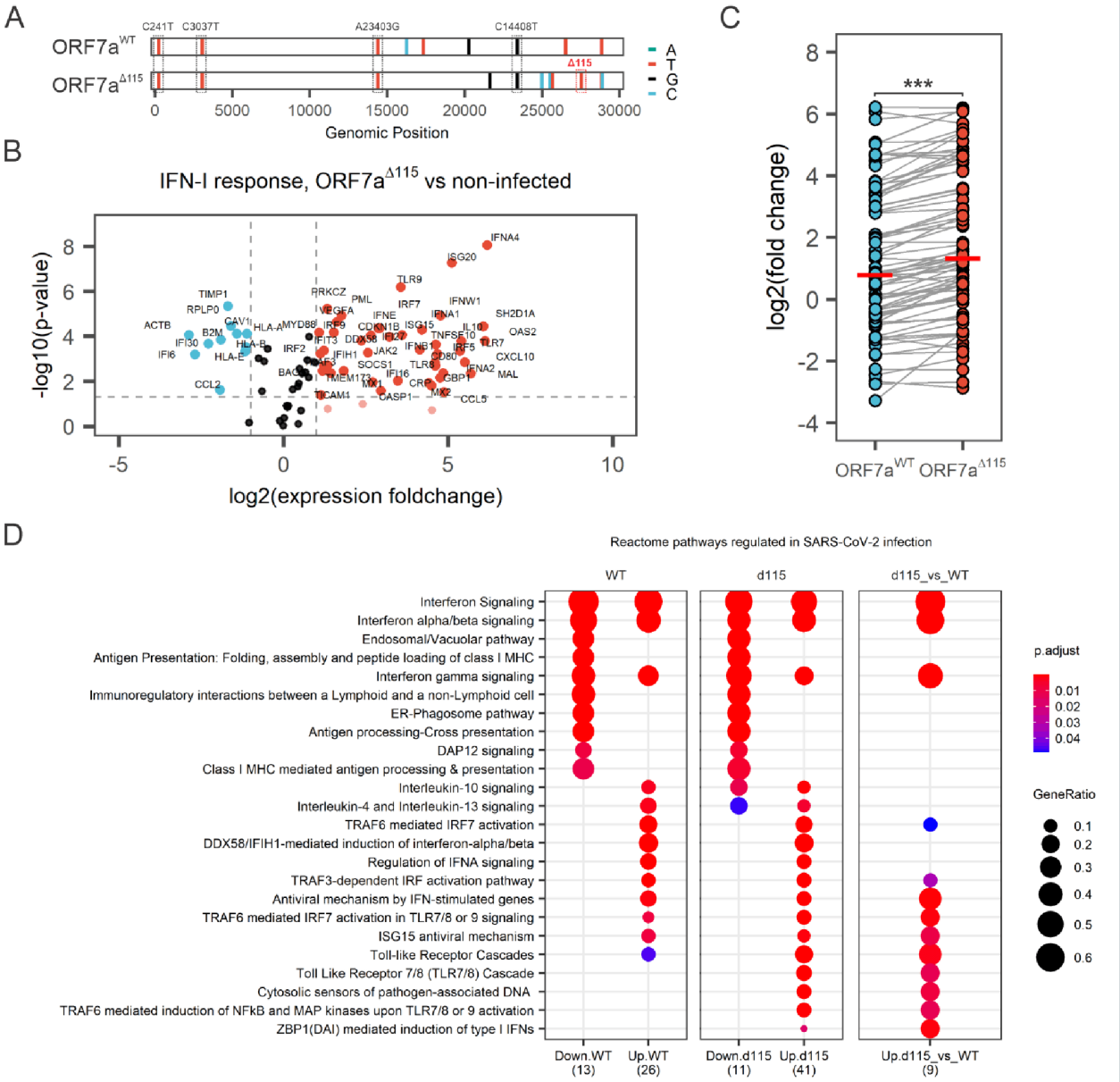
Supplemental Figures



Supplemental Figure S7.1. Phylogenetic and mutational analysis of ORF7a gene sequences. Related to main Figure 7.1. (A) A maximum-likelihood phylogenetic tree indicates that ORF7a$^{\Delta 115}$ strains form a monophyletic clade (red branches). Viral isolate P588 (teal star) was chosen as a "wildtype" control in our experiments. This strain shares a recent common ancestor with the predominant ORF7a$^{\Delta 115}$ clade. SARS-CoV-2 lineages were assigned using pangolin utility (cov-lineages.org).[7] (B) Alignment of unique ORF7a protein sequences with five or more mutations as compared to the SARS-CoV-2 reference strain (Wuhan-Hu-1). Conserved residues are depicted in gray, while mutations and deletions are shown in white. The frequency of each genotype is shown on the right. (C) RT- PCR with primers targeting regions that flank the ORF7a Δ115 deletion sampled in Bozeman, MT. (D) All PCR products in (C) were Sanger sequenced. Representative sequencing chromatograms are shown. Deleted region is shown with dotted red box.



Supplemental Figure S7.2, Mapping TRS-B sites required for ORF7a and ORF7b discontinuous sgRNA transcription. Related to Main Figure 7.2F-H. Direct RNA sequencing data published by Kim et al was used to map TRS-B sites in SARS-CoV-2 ORF7a and ORF7b genes.[11] Vertical lines indicate number of reads spanning the junction between TRS-L and "body" of the gene. Dashed line indicate region deleted in ORF7a$^{\Delta 115}$ strain. The deletion does not remove any of the mapped TRS-B sites suggesting that it has not affected ORF7b sgRNA transcription.



Supplemental Figure S7.3, Related to Main Figure 7.3. IFN-I response and gene set enrichment analysis. (A) Genome sequences of ORF7a^{WT} and ORF7a$^{\Delta115}$ are compared to Wuhan-Hu-1 reference sequence. Colored lines mark differences to reference genome. Mutations that define haplotype of B.1 SARS-CoV-2 lineage are boxed. There are 11 nucleotide positions that differ between WT and Δ115 and are outside ORF7a. Seven of these SNVs change amino acid. In ORF7a^{WT}: S194L in N protein and L1286F in ORF1ab. In ORF7a$^{\Delta115}$: R195I, R203K, G204K in N protein; Q38P and L95I in ORF3a. (B) Volcano plot depicting IFN-I response in HEK 293T-hACE2 cells infected with ORF7a$^{\Delta115}$ SARS-CoV-2 strain at MOI = 0.05. RNA was extracted from infected cells 24 hpi and reverse transcribed. Expression of IFN-I response genes was studied using RT-qPCR array targeting 91 human transcripts (88 targets and 3 references).

Experiment was performed in 3 biological replicates. Dashed lines show regulation (≥2-fold) and statistical significance thresholds (p-value < 0.05). Each dot represents mean (n = 3) normalized expression of a single gene relative to non-infected host. Genes that passed the threshold are labeled. (C) Plot showing distribution of log2(expression fold change) values in ORF7a^{WT} and ORF7a$^{\Delta 115}$ infection vs non-infected control. Each dot represents relative normalized expression for one transcript targeted with the RT-qPCR array. Gray lines connect values for the same transcript in the infection with ORF7a^{WT} and ORF7a$^{\Delta 115}$ variants. Horizontal red bar shows median value. Median ISG activation between two infections was compared using Wilcoxon signed-rank test (***, p-value < 0.001) (D) List of IFN-I response genes that were significantly regulated (p-value < 0.05, ≥ 2-fold change in expression) upon infection with ORF7a$^{\Delta 115}$ and ORF7a^{WT} (vs non-infected control) were analyzed for enrichment of associated Reactome Pathways.[1,55] Gene set enrichment analysis (GSEA) was performed using clusterProfiler package in RStudio.[56] *Down.WT* – genes downregulated during infection with ORF7a^{WT}; *Up.WT* – genes upregulated during infection with ORF7a^{WT}; *Down.Δ115* – Reactome pathways enriched in genes downregulated during infection with ORF7a$^{\Delta 115}$; *Up.Δ115* – genes upregulated during infection with ORF7a$^{\Delta 115}$. *Up.Δ115_vs_WT* – Reactome pathways enriched in genes upregulated during infection with ORF7a$^{\Delta 115}$ vs ORF7a^{WT} (Figure 7.3E).

Materials and Methods

Cell Cultures

HEK293T-hACE2 cells (BEI, NR-52511) were generously provided by Dr. Seth Pincus (Montana State University) and maintained at 37°C and 5% CO2 in Dulbecco's Modified Eagle Medium (DMEM, GIBCO, cat. #12100-061) supplemented with 10% fetal bovine serum (FBS, ATLAS Biologicals, Lot. #F31E18D1), sodium bicarbonate (3.7 g/L), 50 I.U./mL penicillin and 50 µg/mL streptomycin. Vero E6 cells (ATCC® cat. #CRL-1586) were maintained at 37°C and 5% CO2 in Eagle's Modified Eagle Medium (EMEM, ATCC®, cat. #30-2003) supplemented with 10% fetal bovine serum (FBS, ATLAS Biologicals, Lot. #F31E18D1) and 50 I.U./mL penicillin and 50 µg/mL streptomycin. All cell lines tested negative for mycoplasma.

Plasmids

Human-codon optimized dsDNA gene fragments (gBlocks) encoding for SARS-CoV-2 C-terminally FLAG-tagged ORF7aWT and ORF7aΔ115 proteins were synthesized by Integrated DNA Technologies (IDT). Sequence of ORF7a from Wuhan-Hu-1 reference (GenBank MN908947.3) was used for ORF7aWT gBlock. The gene fragments were cloned into the pEGFP-N1 (Clontech cat# 6085-1) backbone to replace the eGFP gene. pLV-mCherry was a gift from Pantelis Tsoulfas (Addgene plasmid # 36084; http://www.addgene.org/36084/; RRID:Addgene_36084).

Antibodies

Western blotting: Rabbit anti-ACTB (Cat: AC026, 1:20,000) antibodies were from ABclonal, Mouse anti-FLAG antibodies (Cat: MA1-91878-1MG, 1:1000) were from

ThermoFisher Scientific, sheep anti-ORF7b antibodies (1:120) were from MRC I PPU, College of Life Sciences, University of Dundee, Scotland; Goat anti-mouse IgG peroxidase-conjugated (Cat: 115-035-003, 1:10,000), Goat anti-rabbit IgG peroxidase-conjugated (Cat: 111-035-003, 1:10,000) and Donkey Anti-Sheep IgG peroxidase-conjugated (Cat: 713-035-147, 1:10,000) antibodies were from Jackson ImmunoResearch. Immunocytochemistry: Mouse anti-FLAG antibodies (Cat: MA1-91878-1MG, 1:200), Goat anti-mouse-AlexaFluor-488 (Cat: A11001, 1:2,000) antibodies were from Invitrogen.

Human clinical sample collection and preparation

Clinical samples were obtained with local IRB approval (protocol #DB033020) and informed consent from patients undergoing testing for SARS-CoV-2 at Bozeman Health Deaconess Hospital. Patients tested for SARS-CoV-2 included both in-patients and out-patients. The latter included individuals who developed symptoms and sought medical care and asymptomatic (at the time of testing) individuals who were exposed to known COVID-19 case and therefore were tested. Nasopharyngeal swabs from patients that tested positive for SARS-CoV-2 were collected in viral transport media. RNA was extracted from all patient samples using QIAamp Viral RNA Mini Kit (QIAGEN) a biosafety level 3 (BSL3) laboratory. All samples were heat-inactivated before removing from BSL3.

Production and titration of coronavirus stocks

The nasopharyngeal swabs in the viral transport media were used to generate viral stocks as previously described.[57] Briefly, 100 ul of the viral transport media was two-fold serially diluted and applied to Vero E6 cells When infected cells showed extensive cytopathic effect, the media was collected and utilized to generate a greater volume second passage in Vero E6 cells.

When CPE was apparent, supernatants were centrifuged 1,000 RCF for 5 minutes to remove cellular debris. Virus titer in clarified supernatants was determined with plaque assay in Vero E6 cells as described.[58] For plaque assay, Vero E6 cells were incubated with viral inoculum at limiting dilutions. Inoculated cells were overlayed with either 0.75% methylcellulose, DMEM supplemented with 2% FBS and 1% pen-strep and incubated for 4 days. Cells were fixed and stained with 0.5% methylene blue in 70% ethanol. Plaques were counted and the overall titer was calculated. Viral stock's identity was confirmed via whole genome sequencing on Oxford Nanopore. All SARS-CoV-2 experiments were performed in a BSL3 laboratory.

Symptom onset data and clinical test results

Suspect cases of COVID-19 were tested in a CLIA lab and instructed to self-quarantine until notified of the RT-qPCR test results. All laboratory confirmed positive cases of COVID-19 were contacted via telephone by local public health nurses to complete contact tracing. During this interview, the nurses collected recorded symptoms, symptom onset date, travel history, contact with other known laboratory confirmed cases, close contacts and activities on the two days before symptom onset up until notification of a positive test. Data collection was conducted as part of a public health response. The study was reviewed by the Montana State University Institutional Review Board (IRB) For the Protection of Human Subjects (FWA 00000165) and was exempt from IRB oversight in accordance with Code of Federal regulations, Part 46, section 101. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Quantitative Reverse Transcription PCR (qRT-PCR)

qRT-PCR was performed using CDC primers (N1 and N2) and probes from the 2019-nCoV RUO Kit (IDT# 10006713). SARS-CoV-2 RNA was quantified using one-step qRT-PCR in ABI 7500 Fast Real-Time PCR System according to CDC protocol (https://www.fda.gov/media/134922/download). In brief, 20 µL reactions included 8.5 µL of Nuclease-free Water, 1.5 µL of Primer and Probe mix (IDT, 10006713), 5 µL of TaqPath 1-Step RT-qPCR Master Mix (ThermoFisher, A15299) and 5 µL of the template. Nuclease-free water was used as negative template control (NTC). Amplification was performed using following program: 25°C for 2 min, 50°C for 15 min, 95°C for 2 min followed by 45 cycles of 95°C for 3 s and 55°C for 30 s. Run data was analyzed in SDS software v1.4 (Applied Biosystems). The NTC showed no amplification throughout the 40 cycles of qPCR.

RT-PCR and SARS-CoV-2 Genome Sequencing

For sequencing 10 mL of RNA from SARS-CoV-2 positive patient sample was reverse transcribed using SuperScript IV (Thermo Fisher Scientific) according to the supplier's protocol. The protocol developed by ARTIC Network was used to generate and sequence amplicon library that covers whole SARS-CoV-2 genome on Oxford Nanopore using ligation sequencing kit (SQK-LSK109) (https://artic. network/ncov-2019). Briefly, two multiplex PCR reactions were performed with primer pools from ARTIC nCoV-2019 V3 Panel (IDT; Table S3) using Q5 High-Fidelity DNA Polymerase (New England Biolabs). Reactions were performed with the following thermocycling conditions: 98ºC for 2min, 30 cycles of 98ºC for 15 s and 65ºC for 5 min. Two resulting amplicon pools for each patient sample were combined and used for library preparation. After end repair (NEB E7546) samples were barcoded using Native Barcoding Expansion Kits EXP-NBD104 and EXP-NBD114 from Oxford Nanopore. A total of 24

barcoded samples were pooled together and Nanopore adaptors were ligated. After clean-up 20 ng of multiplexed library DNA was loaded onto the MinION flowcell for sequencing. A total of 0.3 Gb of raw sequencing data was collected per patient sample. Nuclease-free water was used as a negative control for library preparation and sequencing. Non-specific amplification in negative control was additionally checked using an agarose gel. SARS-CoV-2 Isolate USA-WA1/2020 (BEI, NR-52281) was used as a positive control for genome sequencing.

SARS-CoV-2 Genome Assembly

MinKNOW software was used to basecall raw Nanopore reads in high-accuracy mode. ARTIC bioinformatic pipeline for COVID-19 was used to analyze reads (https://artic.network/ncov-2019). Pipeline included demultiplexing with guppy barcoder and generating consensus sequences with minimap2 and calling single nucleotide variants with nanopolish relative to Wuhan-Hu-1/2019 reference genome.[60-62] 20X coverage was used as a threshold both for consensus and variant calling. Nucleotide positions with less than 20X coverage were masked with Ns in the final consensuses. Consensus sequences were uploaded to GISAID (https://www.gisaid.org/), accession IDs are available in the online supplementary material associated with this manuscript (https://doi.org/10.1016/j.celrep.2021.109197).

RT-PCR and Sanger Sequencing

To verify the ORF7aΔ115 mutation we used RT-PCR with primers that flank the deletion region. Reactions were performed with Q5 High-Fidelity DNA Polymerase in 25 µL volume (New England Biolabs) using following program: 98°C for 2min, 35 cycles of 98°C for 15 s and 65°C for 5 min, 35 cycles. PCR products were analyzed on 1% agarose gels stained with SYBR Safe (Thermo Fisher Scientific). The remaining volume of the reaction was purified using DNA

Clean & Concentrator kit (Zymo Research) and sent to Psomagen for Sanger sequencing. Each PCR product was sequenced with both forward and reverse primers used for PCR. To detect sgRNAs of ORF7a and ORF7b primers annealing to Leader region of SARS-CoV-2 and inside the ORF were used. PCR products were cut out from 1% agarose gel and purified with Zymoclean Gel DNA Recovery Kit. Purified products were sequenced with forward and reverse primers. Sequence was verified by aligning to SARS-CoV-2 genome in UGENE software (Unipro).

Phylogenetic and ORF7a Mutational Analysis

An alignment of 181,003 SARS-CoV-2 genomes was downloaded from the GISAID database at 9:11 AM on 2020-11-11, and sequences without corresponding metadata entries were removed. The nucleotide positions encoding ORF7a in the reference sequence (Wuhan-Hu-1; EP_ISL_402125) were extracted for the remaining 180,971 sequences, and ORF7a sequences with 100% sequence identity were clustered using CD-HIT (with settings: -c 1 -aL 1). One representative was randomly selected from each of the 846 ORF7a sequence clusters and mutations in each nucleotide and translated sequence were determined using the Biostrings package in R. Graphs of these mutations were rendered with the ggplot2 package. To construct a phylogenetic tree of global and Bozeman SARS-CoV-2 sequences, up to ten sequences from each country in the world were selected for each month of the pandemic using the Filter utility in the Augur pipeline (–group-by country year month –sequences-per-group 10).[2,63] This resulted in an alignment of 4,235 GISAID sequences. Then, the Nextclade online tool was used to determine the quality of 73 SARS-CoV-2 genomes (% genome covered) sequenced from Bozeman patients. 55 Bozeman sequences that were determined to be of “Good” quality were

merged with the alignment of GISAID sequences, as well as an alignment of 669 representative SARS-CoV-2 genomes that had non-synonymous mutations in ORF7a. The bat coronavirus RaTG13 (MN996532.2) and the Wuhan- Hu-1 SARS-CoV-2 sequences were also merged, resulting in an alignment of 4,959 full genome sequences, and positions that are hypervariable or prone to sequencing artifacts were masked with a VCF file downloaded from EMBL at 12:45PM on 2020-05- 29. The alignment was trimmed of columns composed of 90% or more gaps using Trimal (-gt 0.9), and IqTree was used to construct a maximum likelihood phylogeny (-m GTR -ninit 2 -n 2 -me 0.05). Branch lengths and internal nodes were rescaled on the resulting tree using TreeTime, and tree was re-rooted to the RaTG13 sequence using the APE package before being visualized with the ggTree package in R. Clades were assigned to SARS-CoV-2 genomes using Nextclade and pangolin (https://cov-lineages.org/) utilities.

Western Blot

293T-hACE2 cells were transfected using Lipofectamine 3000 (Invitrogen, L3000-015) with plasmids encoding for ORF7aWT, ORF7aΔ115 or control pEGFP-N1 plasmid. To detect ORF7b expression cells were infected with ORF7aWT or ORF7aΔ115 SARS-CoV-2 strains at MOI = 0.05. After 24 hours, cells were washed two times with PBS and lysed in RIPA buffer (150 mM sodium chloride, 1% NP-40, 0.5% sodium deoxycholate, 0.2% SDS, 50 mM Tris, pH 8.0) at 4°C for 30 min. The lysates were clarified by centrifugation (10’000 g, 20 min) and stored at −80°C. For western blot lysates were mixed with 6xLaemmli SDS-PAGE buffer and heated at 98°C for 5 min, resolved in 12% SDS-PAGE gel and transferred onto a PVDF (ORF7b) or nitrocellulose (ORF7a) membrane using Mini Trans-Blot Electrophoretic Transfer Cell (Bio-Rad, #1703930). Membranes were blocked and probed with indicated antibodies. The proteins

were visualized using Pierce ECL Western Blotting Substrate (ThermoFisher Scientific, #32106) and exposed to X-Ray film (sc-201696, Santa Cruz Biotech).

Immunocytochemistry

HEK293T-hACE2 were transfected with plasmid DNA using Lipofectamine 3000 (Invitrogen REF: L3000-015) according to manufacturer's instructions. Cells were either transfected with expression vectors for SARS-CoV-2 ORF7aWT, SARS-CoV-2 ORF7aΔ115, or mock transfected. 24 hours after transfection, cells were detached and seeded at 50% confluency on glass coverslips (Fisherbrand, #12-545-80) pre-coated with poly-D-lysine (Cultrex, #3439-100-01) in a 24 well plate. The next day, media was removed, and cells were washed with PBS and fixed with ice-cold 100% methanol on ice for 10 minutes. Half of the methanol was replaced with PBS three times before the methanol solution was discarded and cells were washed three times in PBS (5 minutes each wash). Cells were blocked with 1% BSA (Fisher Scientific BP9703-100) in PBS + 0.1% Tween-20 (PBST) for 30 minutes at room temperature. After blocking, cells were stained with mouse monoclonal anti-Flag antibody (Invitrogen, MA1-91878) (1:200 in 1% BSA in PBST) overnight at 4°C. The next day, cells were washed 3X in PBS (5 min each) and stained with secondary antibody diluted 1:2000 in 1% BSA in PBST for 1 h at RT in the dark (Goat anti-mouse AlexaFluor488; Invitrogen, A11001). After 3x PBS washes nuclei were stained with Hoechst 33342 in PBS for 5 minutes at RT. Coverslips were then mounted with ProLong Gold antifade mounting media (Invitrogen, P36934) on Superfrost Plus microscope slides (#22-037-246). Immunostained cells were imaged using a Leica SP8 confocal microscope.

IFN Inhibition Assay

To generate ISRE reporter system, HEK293T cells were transduced with lentiviruses carrying pGreenFire1-ISRE (Cat. TR016PA-1, SBI) construct and selected with puromycin (1 ug/ml) for 2 days. After selection cells (1x105 cells/well, 48-well plate) were transfected with expression plasmids (250 ng) for SARS-CoV-2 ORF7aWT, SARS-CoV-2 ORF7aΔ115, or control plasmid pLV-mCherry. 16 hours post transfection, cells were treated for 24 h with 5 ng/ml of human IFN-α2b (Cat. rcyc-hifna2b, InvivoGen) according to manufacturer's recommendations. After treatment activation of IFN signaling was measured using Luciferase Assay System (Cat. E1500, Promega) on Cytation 5 (BioTek) plate reader. Assay was performed in six replicates for each condition.

SARS-CoV-2 Replication Assays

HEK293T-hACE2 and Vero E6 cells were seeded in the 48-well plate and infected with ORF7aWT or ORF7aΔ115 SARS-CoV-2 strains at MOI = 0.05 in 50 µL of the cell culture media. Infections were performed in triplicates. Next, 0.5 h post infection (hpi) 500 µL of the fresh cell culture media was added to each well. The supernatants (140 ul) from the infected cells were harvested at the following time points 0.5, 6, 24, 48, 72 and 120 hpi. The RNAs from the supernatants were extracted using QIAamp Viral RNA Mini Kit (QIAGEN) and used as an input in qRT-PCR with CDC N1 and N2 primers and probes as described above. Relative quantification was used (ΔCt method) to calculate viral replication versus 0.5 hpi time point as $2^{-\Delta Ct}$.

The infected cells were harvested 0.5, 6 and 24 hpi and washed with PBS. Total RNA from cells was extracted using RNeasy kit (QIAGEN, Cat No./ID: 74104) with on-column DNase digestion step (QIAGEN, 79254) according to the manufacturers protocol. Viral RNA

was quantified using qRT-PCR CDC N1 primers and probe. Human ACTB endogenous control was quantified with TaqMan assay (ThermoFisher, 4333762T). ΔΔCt method was used to normalize viral RNA level to host RNA and 0.5 hpi. Viral replication was calculated for each replicate as 2-ΔΔCt.

Type I Interferon Response Assay

HEK293T-hACE2 were seeded in the 48-well plate and infected with ORF7aWT or ORF7aΔ115 SARS-CoV-2 strains at MOI = 0.05 in 50 ul of the media. After 30 min on infection the 500 µL of the cell culture media was added to each well. At 24 hpi the cells were washed once with PBS, detached using Trypsin and the total RNA was extracted using RNeasy kit (QIAGEN, Cat No./ID: 74104) with on-column DNase digestion (QIAGEN, Cat No./ID: 79254) according to the manufacturer's protocol. Non-infected cells were processed using the same protocol with mock infection. After extraction, 0,5-1 µg of total RNA was reverse transcribed Reverse using SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) according to the supplier's protocol. The cDNA was diluted 10-fold in water and analyzed with Type I interferon response (SAB Target List) H96 qPCR array. Each reaction contained 2 µL of cDNA, 10 µL of 2X SsoAdvanced Universal SYBR Green Supermix (Bio-Rad) and 8 µL of water. Reactions were performed on QuantStudio 3 Real-Time PCR System instrument with following thermocycling conditions: 95°C for 2min, 40 cycles of 95°C for 5 s and 60°C for 30 s. After amplification, melt curve analysis was performed to examine product specificity. Assay was performed in triplicates for non-infected cells and both viruses. Run data was analyzed in PrimePCR Analysis software (Bio-Rad). Expression of IFN-I response genes was normalized to

geometric mean of 3 reference transcripts (TBP, GAPDH, HPRT1) and non-infected control (ΔΔCt method).

Quantification and Statistical Analysis

All statistical analyses were performed in RStudio v1.2.1335. Moving averages (n = 7) for symptom onset and positive COVID-19 tests data were calculated using geom_ma function from tidyquant package and plotted with ggplot2 in RStudio. qRT-PCR data is shown as mean of three biological replicates (each with three technical replicates) ± standard deviation (SD). Data in ISRE reporter assay was analyzed with one-way ANOVA with Tukey's post hoc pairwise comparisons. Replication of viruses and expression of IFN-I response genes was compared using Student's t test. Medians of IFN-I response between two viruses were compared with Wilcoxon signed-rank test. Significance levels in figures: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ or ns (no significant difference, $p > 0.05$).

CHAPTER EIGHT

INTRINSIC SIGNAL AMPLIFICATION BY TYPE III CRISPR-CAS SYSTEM PROVIDES A SEQUENCE-SPECIFIC SARS-COV-2 DIAGNOSTIC

Contribution of Authors and Co-Authors

Manuscript in Chapter 8

Author: Andrew Santiago-Frangos

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript, wrote the manuscript

Co-Author: Laina N. Hall

Contributions: performed the research

Co-Author: Anna Nemudraia

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Artem Nemudryi

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Pushya Krishna

Contributions: performed the research

Co-Author: Tanner Wiegand

Contributions: designed the research, prepared manuscript figures, performed the research, wrote the manuscript

Co-Author: Royce A. Wilkinson

Contributions: performed the research

Co-Author: Deann T. Snyder

Contributions: acquired samples, performed the research

Co-Author: Jodi F. Hedges

Contributions: acquired samples, performed the research

Co-Author: Calvin Cicha

Contributions: performed the research

Co-Author: Helen Lee

Contributions: performed the research

Co-Author: Ava Graham

Contributions: performed the research

Co-Author: Mark A. Jutila

Contributions: acquired samples, wrote the manuscript

Co-Author: Matthew P. Taylor

Contributions: acquired samples, wrote the manuscript

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

Andrew Santiago-Frangos, Laina N. Hall, Anna Nemudraia, Artem Nemudryi, Pushya Krishna, Tanner Wiegand, Royce A. Wilkinson, Deann T. Snyder, Jodi F. Hedges, Calvin Cicha, Helen H. Lee, Ava Graham, Mark A. Jutila, Matthew P. Taylor, and Blake Wiedenheft

Cell Reports Medicine

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
__X_ Published in a peer-reviewed journal

Cell Press
Volume 2
Issue 6
15 June 2021
DOI: 10.1016/j.xcrm.2021.100319

202

INTRINSIC SIGNAL AMPLIFICATION BY TYPE III CRISPR-CAS SYSTEM PROVIDES A SEQUENCE-SPECIFIC SARS-COV-2 DIAGNOSTIC

Andrew Santiago-Frangos,[1] Laina N. Hall,[1] Anna Nemudraia,[1] Artem Nemudryi,[1] Pushya Krishna,[1] Tanner Wiegand,[1] Royce A. Wilkinson,[1] Deann T. Snyder,[1] Jodi F. Hedges,[1] Calvin Cicha,[1] Helen H. Lee,[1] Ava Graham,[1] Mark A. Jutila,[1] Matthew P. Taylor,[1] and Blake Wiedenheft[1]

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

Abstract

There is an urgent need for inexpensive new technologies that enable fast, reliable, and scalable detection of viruses. Here, we repurpose the type III CRISPR-Cas system for sensitive and sequence-specific detection of SARS-CoV-2. RNA recognition by the type III CRISPR complex triggers Cas10-mediated polymerase activity, which simultaneously generates pyrophosphates, protons, and cyclic oligonucleotides. We show that all three Cas10-polymerase products are detectable using colorimetric or fluorometric readouts. We design ten guide RNAs that target conserved regions of SARS-CoV-2 genomes. Multiplexing improves the sensitivity of amplification-free RNA detection from 10^7 copies/μL for a single guide RNA to 10^6 copies/μL for ten guides. To decrease the limit of detection to levels that are clinically relevant, we developed a two-pot reaction consisting of RT-LAMP followed by T7-transcription and type III CRISPR-based detection. The two-pot reaction has a sensitivity of 200 copies/μL and is completed using patient samples in less than 30 min.

Introduction

Frequent tests and quick results are critical for limiting the spread of severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) and ending the current coronavirus disease 2019 (COVID-19) pandemic.[1,2] qRT-PCR (quantitative reverse transcriptase-polymerase chain reaction) has been the gold standard for viral diagnostics, but this method is slow and requires sophisticated equipment that is expensive to purchase and operate. A national survey of individuals tested using nasal swab-based qRT-PCR found that the average wait times for results was 4.1 days, with 10% of tests taking 310 days.[3] Thus, there is an urgent need for inexpensive new technologies that enable fast, reliable, and scalable detection of viruses.

Recently, loop-mediated isothermal amplification (LAMP)[4] was developed as a sensitive (1–100 copies/mL) point-of-care diagnostic.[5,6] However, LAMP and reverse-transcription LAMP (RT-LAMP) are prone to generating mis-primed or template-independent amplicons that cause false positive readings if they are not coupled to sequence-specific verification of the amplicon.[7-12] For example, the POP7-targeting primers used in the DETECTR CRISPR-based diagnostic lead to excessive nonspecific amplification; however, the sequence-specific verification of the RT-LAMP amplicon by Cas12 yields sensitive and specific detection of POP7 RNA.[12,13] Therefore, platforms that can further verify nucleic acid sequences increase the utility of LAMP and RT-LAMP applications. Type V (Cas12-based) and type VI (Cas13-based) CRISPR systems have been coupled to LAMP or RPA (recombinase polymerase amplification) for sensitive and reliable detection of viral nucleic acids.[13-18] Following isothermal amplification, the RNA-guided Cas12 or Cas13 proteins bind to the amplified target and trigger a non-sequencespecific nuclease activity that cleaves a fluorophore and quencher-labeled DNA or RNA.[14,15,18] Cleavage of the

tether results in an increase in fluorescence that can be detected in 45 min. Furthermore, the collateral nuclease activity of Cas13 has previously been used to cleave and mature an RNA oligo into a linear pseudo-ligand for the RNase Csm6, resulting in a 3.5-fold increase signal sensitivity.[14] While Cas12 and Cas13 detection methods have been optimized over several iterations to be compatible with the isothermal amplification of viral RNA, the ultimate goal is to develop CRISPR-based technologies that are sensitive enough to detect the viral RNA directly, without prior amplification. Recently, Fozouni et al. reported that type IV (Cas13a-based) CRISPR systems can be used for amplification-free detection of SARS-CoV-2 RNA in ~30 min and with the sensitivity of ~100 copies/mL.[19]

Like the type VI (Cas13-based) systems, type III systems also target RNA.[20-24] However, type III systems rely on a unique intrinsic signal amplification mechanism (Figure 8.1). Here, we set out to determine whether intrinsic amplification by type III systems could be used to detect viral RNA without prior amplification. The proof-of-concept presented here demonstrates that the Csm complex from *Thermus thermophilus* (TtCsm) can be programmed to specifically recognize the SARS-CoV-2 genome. SARS-CoV-2, but not SARS-CoV-1 or a panel of other respiratory pathogens, activates the Cas10 polymerase, which generates ~1,000 cyclic nucleotides (e.g., cA4) after binding an RNA target.[25-28] Like all polymerases, nucleotide polymerization by Cas10 also generates protons (H^+) and pyrophosphates (PPi). We demonstrate that each of these three products can be used to detect SARS-CoV-2 RNA using either colorimetric, fluorometric, or both methods simultaneously. The assay can be performed in 1–30 min, depending on the detection method and the concentration of the RNA. When coupled with RTLAMP, the assay can be performed in <30 min and has a limit of detection of ~200

copies/mL. Collectively, this work indicates that the type III systems can be repurposed for fast and sensitive diagnostics.

Results

Sequence-specific Activation of Cas10 Polymerase Yields Three Detectable Products

Sequence-specific recognition of RNA by type III CRISPR systems initiates a signaling cascade (Figure 8.1A).[26-28] RNA binding by the *Thermus thermophilus* (TtCsm) complex triggers a conformational change that activates the Palm domain of the Cas10 subunit, which amplifies the RNA binding signal by converting ATP into ~1,000 cyclic oligoadenylates (e.g., cA_4).[25-28] We hypothesized that the intrinsic signal amplification unique to type III CRISPR systems would boost the sensitivity of direct RNA detection, while maintaining specificity. To test this hypothesis, we expressed and purified the type III-A CRISPR RNA (crRNA)guided surveillance complex from TtCsm with a guide complementary to the N-gene of SARS-CoV-2 ($crRNA_{N1}$) (Figures 8.1B and S8.1; Table S8.1).

The Csm3 subunits, which form the ‘‘backbone’’ of the Csm complex, are nucleases that cleave bound target RNA in 6-nt increments.[22,24,29] The cleaved RNA fragments dissociate and the Csm complex returns to the ‘‘inactive’’ state (i.e., no Cas10-polymerase activity) (Figure 8.1A).[26,30] Thus, in the context of an immune response, the RNase activity of Csm3 moderates Cas10 polymerase activity to limit excess nuclease activation that may otherwise kill the cell.[26,30,31] However, we reasoned that a Csm3 mutation that prevents target RNA degradation would have two related benefits as a diagnostic. First, an RNase-dead Csm complex is expected to stay bound to target RNA longer, which would sustain the Cas10 polymerase activity. Second,

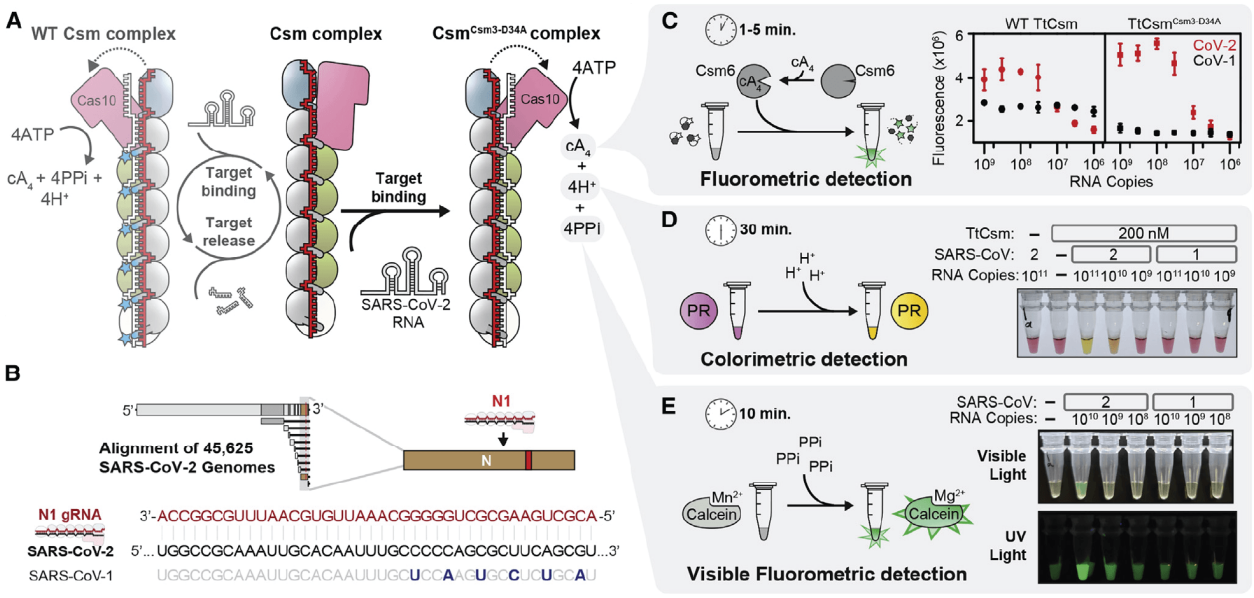


Figure 8.1. Detection of SARS-CoV-2 using the type III CRISPR-Cas system. (A) Schematic of the type III Csm complex from *Thermus thermophilus* (TtCsm). TtCsm complex consists of a crRNA (red) and an unequal stoichiometry of 5 different proteins ($Cas10_1$, pink: $Csm4_1$, blue: $Csm3_6$, gray: $Csm2_4$, green: $Csm5_1$, white). RNA binding activates the Cas10-polymerase, Cas10 DNase, and Csm3 RNase (left, transparent). Target RNA cleavage by Csm3 subunits (blue stars) causes target dissociation and inactivates Cas10. A mutation in the Csm3 subunit ($TtCsm^{Csm3\text{-}D34A}$) renders the complex RNase-dead (right). (B) Schematic of SARS-CoV-2 genome. The N1 CRISPR RNA ($crRNA_{N1}$) targets a sequence on the 3' end of genomic and subgenomic RNAs, which is conserved in 45,625 SARS-CoV-2 genomes. Mismatches between the crRNA-guide and SARS-CoV-1 are highlighted (navy blue). (C) Fluorometric detection of *in vitro* transcribed SARS-CoV-2 (red squares) and SARS-CoV-1 N-gene (black squares). Cyclic tetra-adenylate (cA_4) activates TtCsm6, which is a non-sequence-specific ancillary nuclease (Figure S8.2). Activated TtCsm6 cleaves an RNA tether, which links a fluorophore (star) to a quencher (gray hexagon). $TtCsm^{Csm3\text{-}D34A}$ N1 (right graph) exhibits a LoD 3-fold lower than wild-type TtCsm N1 (left graph) and retains specificity for SARS-CoV-2 RNA. The mean of 3 technical replicates are shown; error bars represent ±1 SD. (D) Colorimetric detection of SARS-CoV-2 RNA by $TtCsm^{Csm3\text{-}D34A}$ N1 complex using a pH-sensitive dye (Phenol Red). Reactions were incubated for 30 min at 60ºC. Technical replicates are shown in Figure S8.3. (E) Visible fluorometric detection of SARS-CoV-2 RNA by $TtCsm^{Csm3\text{-}D34A}$ N1 complex using calcein. Reactions were incubated for up to 1 h at 60ºC. Technical replicates and kinetics are shown in Figure S8.4.

Csm3-mediated cleavage of the target RNA (e.g., SARS-SARCoV-2 RNA) would reduce the target RNA concentration over time and thus limit the sensitivity of the assay. Therefore, we mutated residues in the Csm3 subunit responsible for target RNA cleavage (D34A),[24,29] and

purified the RNase-dead complex (TtCsm$^{Csm3\text{-}D34A}$) (Figure S8.1). To measure the limit of detection (LoD), we added the mutant or wild-type Csm complex to a reaction containing the TtCsm6 nuclease, a fluorescent reporter (i.e., FAM-RNA-Iowa Black FQ), and a titration of RNA corresponding to the N-gene of either SARS-CoV-2 or SARS-CoV-1 (Figures 8.1C and S8.2; Tables S8.2 and S8.3). Single mismatches in the 3' end of target RNAs have been shown to reduce cyclic oligoadenylate production by SepCsm (type III-A) by up to 30-fold as compared to the fully complementary target,[30] while double-nucleotide mutations in the 3' end of target RNAs ablated cyclic oligoadenylate production by SsoCsm (type III-D).[26] Furthermore, single- and double-nucleotide complementarity between the crRNA 5' handle and the target 3' anti-tag significantly decreased the allosteric activation of Cas10 and cyclic oligoadenylate production by SthCsm (type III-A).[32] Single mismatches in the target RNA have been shown to result in 10-fold lower amounts of cyclic oligoadenylate production by other Csm complexes.[30] Using fluorometric detection, both the mutant and the wild-type Csm complex could detect the SARS-CoV-2 RNA at concentrations >10^8 copies per reaction, and neither complex cross-reacted with the SARS-CoV-1 RNA at the highest concentrations tested. The RNase-dead TtCsm complex was roughly 3-fold more sensitive than the wild type, with an LoD of ~10^7 copies per reaction.

In addition to fluorometric detection, we developed a colorimetric RNA detection method that uses a pH change that occurs during nucleotide polymerization (Figures 8.1D and S8.3). Specific recognition of SARS-CoV-2 by RNase-dead TtCsm complex activates Cas10. Cas10 polymerizes ATP,[25-28] releasing one proton per incorporated nucleotide. Cas10-generated protons acidify the solution and change the color of a pH indicator (i.e., Phenol Red) from fuchsia through orange (10^{10} RNA copies) to yellow (10^{11} RNA copies). Similarly, we developed a

visible fluorometric detection method that relies on the sequestration of metallic ions by pyrophosphate. The metal indicator calcein is initially quenched by bound Mn^{2+} ions.[33] In addition to the cyclic oligoadenylates and protons, Cas10 polymerase generates one pyrophosphate per ATP polymerized. Pyrophosphate forms an insoluble precipitate with Mn^{2+}, which unquenches calcein. Free calcein is then bound by excess Mg^{2+}, forming a fluorescent complex that can be seen by eye or with a UV lamp in <10 min (Figures 8.1E and S8.4).

Csm-based Direct Detection of SARS-CoV-2 RNA in Patient Samples

The LoD using $crRNA_{N1}$ is between 10^7 and 10^8 copies of IVT RNA per microliter, which is insufficient to be clinically relevant (Figures 8.1C and 8.2C). To identify other guides that may outperform or complement the activity of $crRNA_{N1}$, we aligned 45,641 SARS-CoV-2 genomes available from GISAID.[34] These alignments were used to select guides based on four key criteria. First, each target sequence had to be >99% identical among the available SARS-CoV-2 genomes. Second, complementarity between the target and the crRNA was not allowed to extend beyond the spacer sequence (guide) and into repeat derived portions of the crRNA that have been shown to suppress Cas10 activity.[27] Third, we targeted regions of SARS-CoV-2 that were different by at least 2 nt in SARS-CoV-1 and MERS-CoV. Fourth, the list of target sequences was pruned to remove guides with similarity to human mRNA sequences or common oral and respiratory pathogen sequences (E value <1,000). Finally, we focused on target sequences located 3' of the ORF3a gene, which are present on both the viral genome and subgenomic RNAs generated during infection. In total, we designed crRNAs targeting 10 different locations on the SARS-CoV-2 genome (Figures 8.2A and S8.1; Table S8.1). To determine how each of these guides perform, we measured sequence-specific detection of RNA

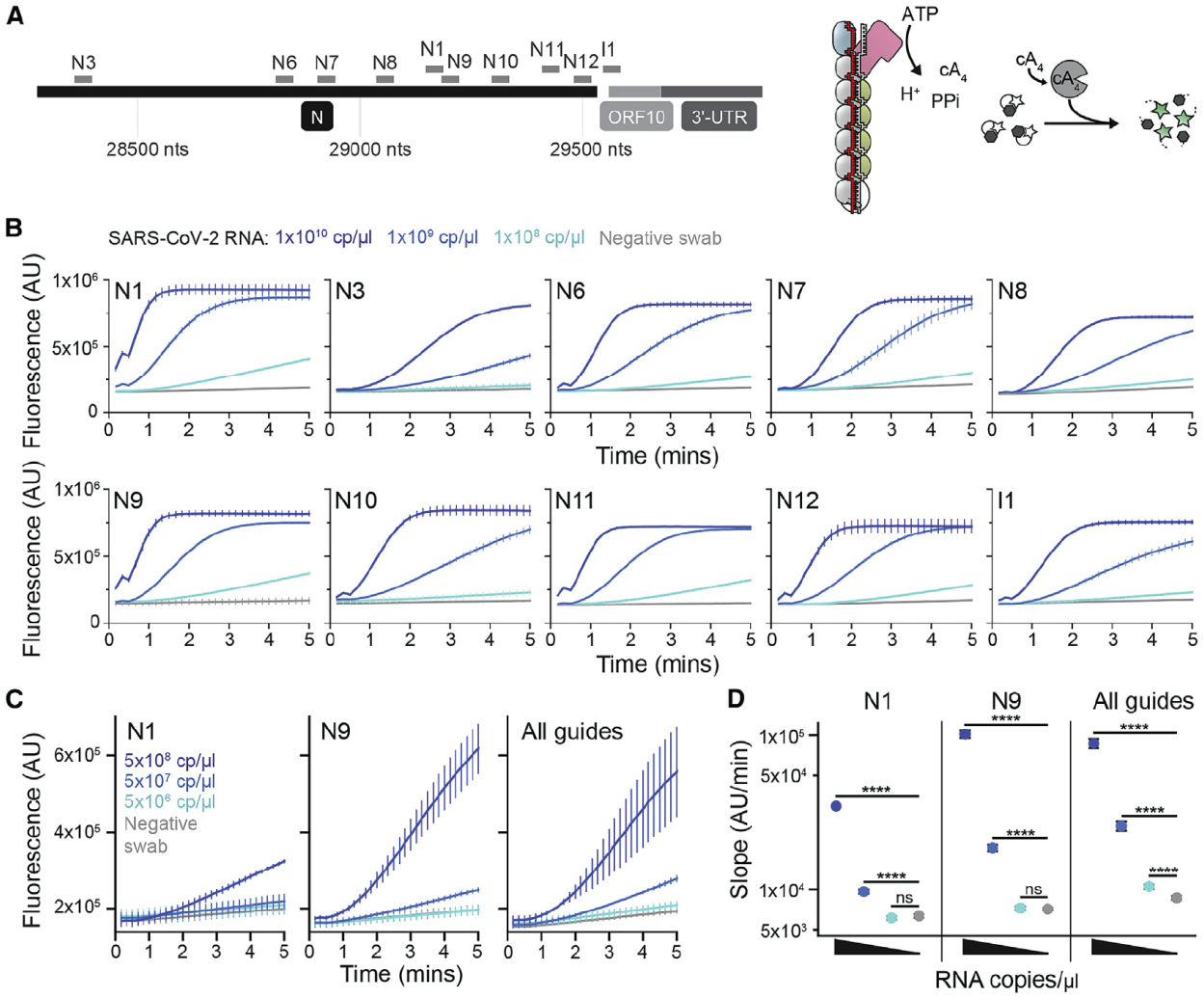


Figure 8.2. Screening and multiplexing crRNA guides. (A) Regions of SARS-CoV-2 genome targeted by each of the 10 guides, and schematic of RNA reporter-based assay (right) used to test guides. (B) Detection of SARS-CoV-2 IVT RNA, spiked into RNA extracted from patients lacking a SARS-CoV-2 infection, by 10 different TtCsm$^{Csm3\text{-}D34A}$ complexes (25 nM), via a reporter RNA-based assay (Figure 8.1C; Tables S8.1 and S8.3). Means and SDs of 2 technical replicates are shown. (C) Direct detection of SARS-CoV-2 genomes in RNA extracted from patient samples by 25 nM TtCsm$^{Csm3\text{-}D34A}$ N1 or N9, or a mixture of 10 different TtCsm$^{Csm3\text{-}D34A}$ complexes, each at 2.5 nM, via a reporter RNA-based assay. RNA extracted from a patient with a high viral load (5 3 10^8 copies/mL as determined by qRT- PCR) was diluted into RNA extracted from patients lacking a SARS-CoV-2 infection. Means and SDs of 3 technical replicates are shown. (D) Slopes of increasing fluorescence, measured in (C), were calculated by simple linear regression. The calculated slope and ±95% confidence intervals are shown. Positive RNA slopes were compared to the negative swab RNA slope by an F-test: ****$p < 0.0001$, ns, not significantly higher than negative swab RNA control.

using a fluorometric reporter assay (i.e., FAM-RNA-Iowa Black FQ) (Figure 8.2B). Most of the crRNAs provide similar sensitivity; however, $crRNA_{N1}$ and $crRNA_{N9}$ generated significantly more signal than the next best complex tested ($p < 0.0001$). We then tested $crRNA_{N1}$ and $crRNA_{N9}$ on RNA isolated from the nasal swab of an infected patient (Figure 8.2C). Both $crRNA_{N1}$- and $crRNA_{N9}$-guided complexes generate a similar signal for either IVT RNA or RNA isolated from a SARS-CoV-2^{+} patient (i.e., 2- to 3-fold increase in signal by 5 min, relative to the first time point). The LoD for $crRNA_{N1}$ and $crRNA_{N9}$ is 10^7 copies/mL (Figure 8.2D) ($p < 0.0001$).

Fozouni et al. recently showed that multiplexing Cas13 (i.e., combining multiple guides into a single reaction) improves the sensitivity of SARS CoV-2 detection.[19] We reasoned that similar benefits may be possible for Csm-based detection. To test this idea, we combined 10 of the guides (2.5 nM each) into a single multiplexed reaction. Multiplexing 10 guides improves the sensitivity of TtCsm-mediated detection of SARS-CoV-2 RNA isolated from the nasal swab of a positive patient by ~10 times (Figures 8.2C and 8.2D). However, the sensitivity of direct detection appears to increase additively with the number of TtCsm complexes, which precludes the direct detection of RNA at concentrations that are clinically relevant.

Testing Clinical Samples for SARS-CoV-2 using RT-LAMP and T7-Csm

Csm-based detection is currently not sensitive enough to directly detect SARS-CoV-2 in all of the patients capable of spreading the infection, which requires an LoD of 10^3 RNA copies/mL (Figure 8.2C).[1,2,35,36] To decrease the LoD of a type III CRISPR-based diagnostic to ~10^3 RNA copies/mL, we incorporated an upstream nucleic acid amplification technique (Figure 8.3A). SARS-CoV-2 genomic RNA is reverse transcribed (RT) into DNA, which is then amplified by

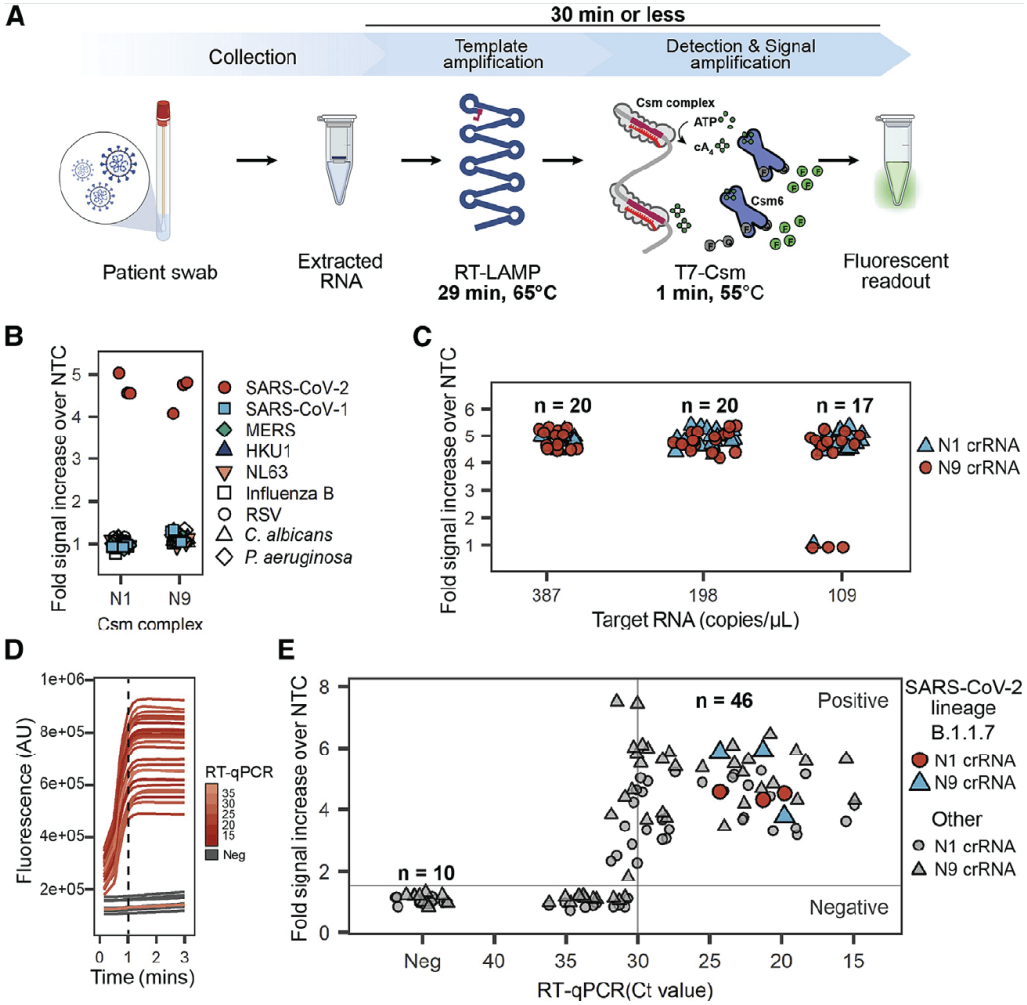


Figure 8.3. CRISPR-Csm-based detection of SARS-CoV-2 in clinical nasopharyngeal swab samples. (A) Schematic of RT-LAMP-T7-Csm based detection. The viral RNA is reverse transcribed, and the resulting DNA is amplified in an RT-LAMP reaction to produce transcription templates for T7 RNA polymerase, in one pot. An aliquot of the RT-LAMP reaction (29 min) is then mixed with the T7-Csm reaction (1 min). (B) RT-LAMP-T7-Csm is specific. Neither $crRNA_{N1}$ nor $crRNA_{N9}$ cross-react with other coronaviruses, or other common human oral pathogens or flora. Detection of SARS-CoV-2 by both crRNAs is rapid, 1 min, and robust, 4- to 5-fold increase in signal over no-template control (NTC). Three technical replicates are shown. (C) RT-LAMP-T7-Csm is sensitive. $TtCsm^{Csm3\text{-}D34A}$ complexes loaded with either $crRNA_{N1}$ or $crRNA_{N9}$ have LoDs of 198 copies/mL (20/20 technical replicates); 20 technical replicates are shown. (D) Kinetics of fluorescence signal increase in T7-Csm reactions. SARS-CoV-2^{+} patient samples are observed to have a ~2-fold increase in signal over NTC by 10 s ($crRNA_{N1}$ median = 1.8, $crRNA_{N9}$ median = 2.3), which rapidly increases to a 4-5-fold increase in signal over an NTC reaction by 1 min ($crRNA_{N1}$ median = 4.2, $crRNA_{N9}$ median = 5.5). A subset of traces is shown for clarity (Figure S8.5). (E) Nasopharyngeal swabs from 56 individuals were tested with qRT-PCR (x-axis) and RT- LAMP-T7-Csm (y-axis). Swabs with Ct values <40 for both N1 and N2 CDC diagnostic primers are considered positive for SARS-CoV-

2 RNA. RT-LAMP-T7-Csm reliably identifies patient samples with a Ct < 30.7 (200–100 RNA copies/mL) as positive for SARS-CoV-2 (Figure S8.6). The B.1.1.7 variant is positively identified by both $crRNA_{N1}$ and $crRNA_{N9}$. Data are shown as fold change in fluorescence as compared to NTC reaction. Single technical replicates are shown for qRT-PCR and RT-LAMP-T7- Csm experiments with either N1 or N9 crRNAs.

LAMP using primers that flank regions of the SARS-CoV-2 genome targeted by $crRNA_{N1}$ and $crRNA_{N9}$. One of the LAMP primers incorporates a T7 promoter into the amplified DNA, which is then used for *in vitro* transcription (T7) and detected by TtCsm (Figures 8.3A and S8.5; Table S8.4).

To confirm the specificity of TtCsm-based detection, we tested SARS-CoV-2 alongside a panel of eight other oral and respiratory pathogens, including coronaviruses SARS-CoV-1, Middle East respiratory syndrome coronavirus (MERS-CoV), human coronavirus HKU1, and human coronavirus NL63 (Figure 8.3B). These samples resulted in a background signal similar to the no-template control (NTC) (Figure 8.3B). In contrast, SARS-CoV-2 RNA results in a 4- to 5-fold increase in signal.

To determine the LoD of RT-LAMP-T7-Csm, we tested 20 replicates of 2-fold serial dilutions ranging from ~100-400 copies/mL SARS-CoV-2 RNA (Figure 8.3C). The LoD of RT-LAMP-T7-Csm is 198 copies/mL SARS-CoV-2 RNA (20/20 replicates), in an assay that relies on a 29-min RT-LAMP step, followed by a 1-min T7Csm fluorometric detection reaction (Figure 8.3D; Table 8.1).

To further validate this method, we next tested RNA extracted from 56 nasopharyngeal swab samples taken from patients who had previously been tested using qRT-PCR. Of the 56 samples tested, 46 were positive for SARS-CoV-2 and 10 were negative by qRT-PCR (Figure 8.3E). Using two different crRNA guides, we demonstrate that the type III CRISPR system has a

Test	Time to result (min)	LoD (copies/μL)	Positive agreement (%)		Negative agreement (%)
			2x LoD	All Samples Tested	All Samples Tested
DETECTR	45	20	95.0%	95%	100.0%
SHERLOCK	50	6.75	100.0%	100.0%	100.0%
RT-LAMP-T7-Csm	30	100	100%	74%	100.0%

Table 8.1. Comparison of clinical parameters of RT-LAMP-T7-Csm and FDA EUA-approved protocols of DETECTR and SHERLOCK diagnostics.[37,38]

specificity (negative predictive agreement) of 100%, as well as a positive predictive agreement of 100% for nasopharyngeal swab samples with 100–200 copies/mL SARS-CoV-2 RNA as determined by qRT-PCR (Figures 8.3E and S8.5). Whole-genome sequencing revealed that three of the patient samples used here belonged to the B.1.1.7. lineage. These genome sequences have been deposited in GISAID accession IDs: EPI_ISL_1081321, EPI-ISL_1081322, EPI_ISL_1081323).[34] Importantly, the B.1.1.7. variants were positively identified by RT-LAMP-T7-Csm with both N1 and N9 crRNA guides (Figure 8.3E; N1, red circles; N9, blue triangles).

Discussion

Collectively, this work demonstrates that sequence-specific detection of viral RNA by the type III CRISPR-Cas complex triggers the synthesis of cA4, pyrophosphate, and protons, each of which are detectable within 1–30 min, using colorimetric or fluorometric methods (Figure 8.1). Coupling RT-LAMP to T7 transcription and Csm-based detection creates a rapid (<30 min) testing protocol with attomolar sensitivity and high specificity. Moreover, we show that this protocol is capable of detecting the B.1.1.7 SARS-CoV-2 variant, and we anticipate that the guide design criteria described here could be applied to the specific detection of other pathogens. Furthermore, the RT-LAMP and T7-Csm reactions occur in similar buffers and at similar temperatures; therefore, it is possible that additional modifications that either increase the

efficiency of RT-LAMP at 55ºC (i.e., redesigning primer sets), or that increase the efficiency of T7-Csm at 65ºC (i.e., a more thermostable RNA polymerase), could enable an integrated one-pot diagnostic.

While the LoD for RT-LAMP-T7-Csm is 1–1.5 orders of magnitude higher than FDA Emergency Use Authorization (EUA)-approved DETECTR and SHERLOCK assays (20 and 6.75 copies/mL, respectively), the time to result is ~30% faster, which would translate to a higher throughput of samples analyzed per unit of machine run time. Sensitivity was strategically sacrificed for time in this proof-of-concept application for optimal detection of infectious SARS-CoV-2 patients, which requires an LoD of 1,000 copies/mL.[1,2,35,36] However, an increase in the length of the RT-LAMP step (i.e., to 40 min, as is used in SHERLOCK), would likely increase the sensitivity. Additional modifications that could further increase sensitivity include the optimization of RT-LAMP primer sets, a screen for more active type III Csm or Cmr surveillance complexes, or a screen for CARF-fusion effectors that do not cleave their cyclic oligoadenylate ligands.[39] Alternatively, more sensitive methods of measuring Cas10-generated protons, such as ion-sensitive field-effector transistors used in label-free sequencing by synthesis technologies, could decrease the LoD.

Type V (Cas12-based) and type VI (Cas13-based) CRISPRbased diagnostics hinge completely on their collateral nuclease activity.[13-17,40] The Cas10 subunits of Csm complexes possess an analogous nuclease activity triggered upon target binding.[21,29,41-43] However, here, we focused on the NTP polymerase activity of the activated Csm complex, which enables readouts based on several different chemistries (Figure 8.1). Furthermore, there is a rich resource of naturally occurring downstream effector proteins that have evolved to be activated by the Cas10

polymerized NTP products, which possess a wide range of enzymatic activities. These effectors include RNases, DNases, putative proteases, nitrilases, adenosine deaminases, and adenylate cyclases.[44] Future efforts are aimed at incorporating complementary effectors with the goal of reducing the LoD for the direct detection of RNA in clinical samples.

Limitations of Study

This study demonstrates a proof-of-concept for a sensitive, specific, and rapid diagnostic based on the type III CRISPR system. The current implementation requires pre-amplification using RTLAMP, followed by T7-transcription and type III detection (T3D). Clinical implementation of this approach will benefit from changes that eliminate the need for RT-LAMP and T7 amplification, improvements that limit sample handling, as well as a better understanding of the base pairing requirements that activate the Cas10 polymerase.[16,19,26,30,32] Improvements will benefit from screening additional guides, testing other type III complexes for accelerated polymerase activity, and incorporating other ancillary nucleases or other effectors in a way that boosts the sensitivity and reduces the time to result.

Acknowledgements

We are grateful to members of Bozeman Health who provided deidentified patient samples, specifically Christopher Nero, Douglas Smoot, Winfield Wallace, Cayley Faurot-Daniels, Verena Lawrence, and Melissa Blauvelt. We are also grateful to Michelle Flenniken, Katie Daughenbaugh, Diane Bimczok, and other members of the COVID task force at Montana State University (MSU) for assistance establishing the COVID testing center. A.S.-F. is a postdoctoral fellow of the Life Science Research Foundation, which is supported by the Simons

Foundation. A.S.-F. is supported by the Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund. Research in the Wiedenheft lab and the biosafety level 3 facility is supported by the Office of the Vice President for Research at Montana State University, a generous donation from the Rosolowsky family, and a sponsored research agreement from VIRIS Detection Systems. Some graphical assets were created with BioRender.com.

B.W. is the founder of SurGene and VIRIS Detection Systems. B.W., A.S.-F., A. Nemudraia, and A. Nemudryi are inventors on patent applications related to CRISPR-Cas systems and applications thereof.

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. One or more of the authors of this paper received support from a program designed to increase minority representation in science. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

References

1 Larremore, D. B. *et al.* Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv* **7**, doi:10.1126/sciadv.abd5393 (2021).

2 Paltiel, A. D., Zheng, A. & Walensky, R. P. Assessment of SARS-CoV-2 Screening Strategies to Permit the Safe Reopening of College Campuses in the United States. *JAMA Netw Open* **3**, e2016818, doi:10.1001/jamanetworkopen.2020.16818 (2020).

3 Lazer, D. S., M.; Perlis, R.H.; Ognyanova, K.; Baum, M.A.; Quintana, A.; Druckman, J.; Volpe, J.D.; Chwe, H.; Simonson, M. State of the Nation: A 50-State COVID-19 Survey: Report #8: Failing the Test: Waiting Times for COVID Diagnostic Tests Across the U.S., (World Health Organization, 2020).

4 Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res* **28**, E63, doi:10.1093/nar/28.12.e63 (2000).

5 Dao Thi, V. L. *et al.* A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci Transl Med* **12**, doi:10.1126/scitranslmed.abc7075 (2020).

6 Zhang, Y. *et al.* Enhancing colorimetric loop-mediated isothermal amplification speed and sensitivity with guanidine chloride. *Biotechniques* **69**, 178-185, doi:10.2144/btn-2020-0078 (2020).

7 Meagher, R. J., Priye, A., Light, Y. K., Huang, C. & Wang, E. Impact of primer dimers and self-amplifying hairpins on reverse transcription loop-mediated isothermal amplification detection of viral RNA. *Analyst* **143**, 1924-1933, doi:10.1039/c7an01897e (2018).

8 Hardinge, P. & Murray, J. A. H. Reduced False Positives and Improved Reporting of Loop-Mediated Isothermal Amplification using Quenched Fluorescent Primers. *Sci Rep* **9**, 7400, doi:10.1038/s41598-019-43817-z (2019).

9 Gao, X., Sun, B. & Guan, Y. Pullulan reduces the non-specific amplification of loop-mediated isothermal amplification (LAMP). *Anal Bioanal Chem* **411**, 1211-1218, doi:10.1007/s00216-018-1552-2 (2019).

10 Rolando, J. C., Jue, E., Barlow, J. T. & Ismagilov, R. F. Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification. *Nucleic Acids Res* **48**, e42, doi:10.1093/nar/gkaa099 (2020).

11 Zou, Y., Mason, M. G. & Botella, J. R. Evaluation and improvement of isothermal amplification methods for point-of-need plant disease diagnostics. *PLoS One* **15**, e0235216, doi:10.1371/journal.pone.0235216 (2020).

12 Ooi, K. H. *et al.* An engineered CRISPR-Cas12a variant and DNA-RNA hybrid guides enable robust and rapid COVID-19 testing. *Nat Commun* **12**, 1739, doi:10.1038/s41467-021-21996-6 (2021).

13 Broughton, J. P. *et al.* CRISPR-Cas12-based detection of SARS-CoV-2. *Nat Biotechnol* **38**, 870-874, doi:10.1038/s41587-020-0513-4 (2020).

14 Gootenberg, J. S. *et al.* Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439-444, doi:10.1126/science.aaq0179 (2018).

15 Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436-439, doi:10.1126/science.aar6245 (2018).

16 Joung, J. *et al.* Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing. *N Engl J Med* **383**, 1492-1494, doi:10.1056/NEJMc2026172 (2020).

17 Gootenberg, J. S. *et al.* Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438-442, doi:10.1126/science.aam9321 (2017).

18 East-Seletsky, A. *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270-273, doi:10.1038/nature19802 (2016).

19 Fozouni, P. *et al.* Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **184**, 323-333 e329, doi:10.1016/j.cell.2020.12.001 (2021).

20 Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956, doi:10.1016/j.cell.2009.07.040 (2009).

21 Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, C. & Siksnys, V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Mol Cell* **62**, 295-306, doi:10.1016/j.molcel.2016.03.024 (2016).

22 Samai, P. *et al.* Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* **161**, 1164-1174, doi:10.1016/j.cell.2015.04.027 (2015).

23 Staals, R. H. *et al.* RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. *Mol Cell* **56**, 518-530, doi:10.1016/j.molcel.2014.10.005 (2014).

24 Tamulaitis, G. *et al.* Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. *Mol Cell* **56**, 506-517, doi:10.1016/j.molcel.2014.09.027 (2014).

25 Jia, N., Jones, R., Sukenick, G. & Patel, D. J. Second Messenger cA4 Formation within the Composite Csm1 Palm Pocket of Type III-A CRISPR-Cas Csm Complex and Its Release Path. *Mol Cell* **75**, 933-943 e936, doi:10.1016/j.molcel.2019.06.013 (2019).

26 Rouillon, C., Athukoralage, J. S., Graham, S., Gruschow, S. & White, M. F. Control of cyclic oligoadenylate synthesis in a type III CRISPR system. *Elife* **7**, doi:10.7554/eLife.36734 (2018).

27 Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609, doi:10.1126/science.aao0100 (2017).

28 Niewoehner, O. *et al.* Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548, doi:10.1038/nature23467 (2017).

29 Liu, T. Y., Iavarone, A. T. & Doudna, J. A. RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. *PLoS One* **12**, e0170552, doi:10.1371/journal.pone.0170552 (2017).

30 Nasef, M. *et al.* Regulation of cyclic oligoadenylate synthesis by the Staphylococcus epidermidis Cas10-Csm complex. *RNA* **25**, 948-962, doi:10.1261/rna.070417.119 (2019).

31 Athukoralage, J. S. *et al.* The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling. *Elife* **9**, doi:10.7554/eLife.55852 (2020).

32 You, L. *et al.* Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. *Cell* **176**, 239-253 e216, doi:10.1016/j.cell.2018.10.052 (2019).

33 Tomita, N., Mori, Y., Kanda, H. & Notomi, T. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nat Protoc* **3**, 877-882, doi:10.1038/nprot.2008.57 (2008).

34 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

35 La Scola, B. *et al.* Viral RNA load as determined by cell culture as a management tool for discharge of SARS-CoV-2 patients from infectious disease wards. *Eur J Clin Microbiol Infect Dis* **39**, 1059-1061, doi:10.1007/s10096-020-03913-9 (2020).

36 Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465-469, doi:10.1038/s41586-020-2196-x (2020).

37 Biosciences, M. SARS-CoV-2 RNA DETECTR Reagent Kit. (https://www.fda.gov/media/141765/download, 2020).

38 Biosciences, S. Instructions for use: SherlockTM CRISPR SARS-CoV-2 kit. (https://www.fda.gov/media/137746/download, 2020).

39 Jia, N., Jones, R., Yang, G., Ouerfelli, O. & Patel, D. J. CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity. *Mol Cell* **75**, 944-956 e946, doi:10.1016/j.molcel.2019.06.014 (2019).

40 Yan, W. X. *et al.* Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88-91, doi:10.1126/science.aav7271 (2019).

41 Sofos, N. *et al.* Structures of the Cmr-beta Complex Reveal the Regulation of the Immunity Mechanism of Type III-B CRISPR-Cas. *Mol Cell* **79**, 741-757 e747, doi:10.1016/j.molcel.2020.07.008 (2020).

42 Elmore, J. R. *et al.* Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev* **30**, 447-459, doi:10.1101/gad.272153.115 (2016).

43 Zhang, J., Graham, S., Tello, A., Liu, H. & White, M. F. Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. *Nucleic Acids Res* **44**, 1789-1799, doi:10.1093/nar/gkw020 (2016).

44 Makarova, K. S. *et al.* Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. *Nucleic Acids Res* **48**, 8828-8847, doi:10.1093/nar/gkaa635 (2020).

45 Liu, T. Y., Liu, J. J., Aditham, A. J., Nogales, E. & Doudna, J. A. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. *Nat Commun* **10**, 3001, doi:10.1038/s41467-019-10780-2 (2019).

46 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518, doi:10.1093/nar/gki198 (2005).

47 Nemudryi, A. *et al.* SARS-CoV-2 genomic surveillance identifies naturally occurring truncation of ORF7a that limits immune suppression. *Cell Rep* **35**, 109197, doi:10.1016/j.celrep.2021.109197 (2021).

48 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8, doi:10.1186/s13059-018-1618-7 (2019).

49 Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*, doi:10.1101/2020.09.04.283077 (2020).

50 Loman, N. J. R., W.; Rambaut, A. NCoV-2019 Novel Coronavirus Bioinformatics Protocol. (Artic Network, 2020).

51 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).

Supplemental Figures

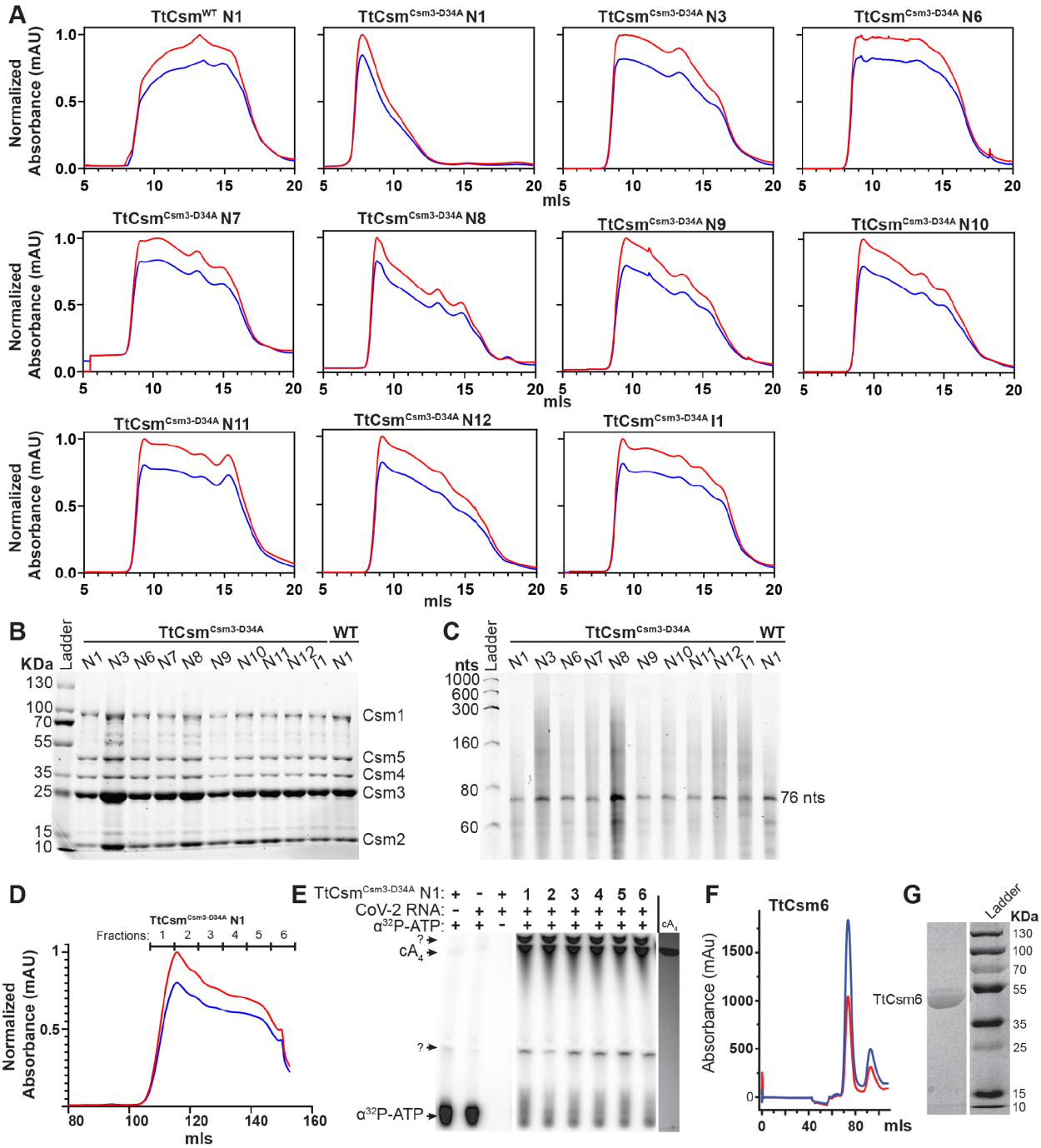


Figure S8.1. Purification of TtCsm complexes and TtCsm6 nuclease. Related to Figures 8.1-8.3. (A) Size exclusion chromatography (SEC) profiles of TtCsm WT and TtCsm Csm3-D34A complexes loaded with different crRNA guides. SEC performed using a Superose 6 Increase 10/300 GL size-exclusion column (Cytiva). Normalized absorbance (mAU) was measured at

260nm (red) and 280nm (blue). Fractions 9 up to 16 of each SEC purification were collected, concentrated, and stored a t -80°C. (B) The fractions were combined, concentrated and run on an SDS-PAGE. All five Csm proteins are present and the intensities of each band correspond with our understanding of the protein stoichiometry of the assembled TtCsm complex. (C) RNA was isolated from the pooled and concentrated SEC fractions. Denaturing urea polyacrylamide gel of nucleic acids associated with each TtCsm complex. The full-length crRNA intermediate is expected to be 76 nucleotides (nts) long. (D) SEC profile of TtCsmCsm3-D34A N1 complex. Six successive fractions, representing the entire peak, were collected concentrated and stored separately. (E) Sequence-specific activation of Cas10 was estimated for each of the six fractions in panel D. ^{32}P-ATP polymerization was measured using thin-layer chromatography (TLC). 500nM TtCsm$^{Csm3\text{-}D34A}$ N1 complex was incubated with 10^{10} copies of target RNA, 50 μM ATP and 10 nM α-^{32}P-ATP, a t 60°C for 1 hour. Nucleic acids were phenol-chloroform extracted from each reaction and spotted on a silica gel TLC plate coated with fluorescent indicator F254, developed in solvent (0.2 M ammonium bicarbonate pH 9.3, 70% ethanol, 30% water). An unlabeled cA4 standard (Axxora) was run on the same TLC plate, in a parallel lane, and was visualized by illumination with a handheld shortwave (254 nm) UV lamp (Analytik Jena) and a Galaxy S9 phone (Samsung). One of the two major ^{32}P-labelled products generated by target RNA-bound TtCsm complex migrates similarly to the cA_4 standard. All TtCsm complexes polymerize similar amounts of α-^{32}P-ATP (bottom band) into similar ratios of similarly migrating products (top bands). (F) SEC profile of TtCsm6 ancillary nuclease purified on a Superdex 200 26/600 size-exclusion column (Cytiva). (G) SDS-PAGE gel image of purified TtCsm6.

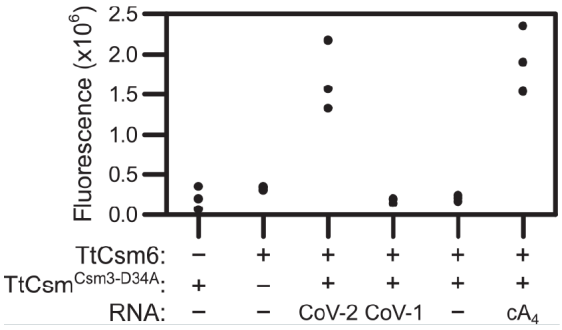


Figure S8.2. Csm-mediated synthesis of cA_4 is necessary for activation of the non-sequence specific nuclease Csm6. Related to Figure 8.1. Detection of invitro transcribed SARS-CoV-2 RNA using the type III surveillance complex TtCsm$^{Csm3\text{-}D34A}$ and the ancillary nuclease, TtCsm6. TtCsm$^{Csm3\text{-}D34A}$, TtCsm6 and SARS-CoV-2 RNA, but not SARS-CoV-1 RNA, are all required to trigger cleavage of the fluorescent RNA reporter. Addition of cA^4, bypasses the signaling pathway and activates TtCsm6 directly, cleaving the reporter RNA in the absence of SARS-CoV-2 RNA (last lane). These reactions were performed using10^8 copies of *in vitro* transcribed SARS-CoV-1 or -2 RNA spiked into nasopharyngeal swab clinical matrix. Results from three technical replicates are shown. Results presented in the final column (cA4) demonstrate that the florescent signal is generated by cA^4-dependent activation of Csm6.

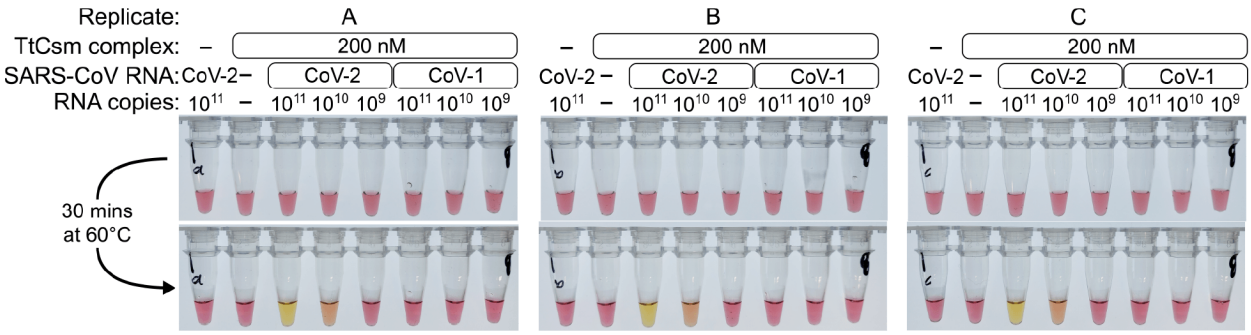


Figure S8.3. Colorimetric detection of a specific RNA sequence. Related to Figure 8.1. TtCsm$^{Csm3\text{-}D34A}$ was incubated with either RNA target, in the presence of ATP, for 30 minutes at 60°C. Specific RNA recognition, and associated Cas10-mediated ATP polymerization, causes acidification of the solution, and changes the color of Phenol Red pH indicator from fuchsia, through orange, to yellow. Three technical replicates (A, B and C) are shown.

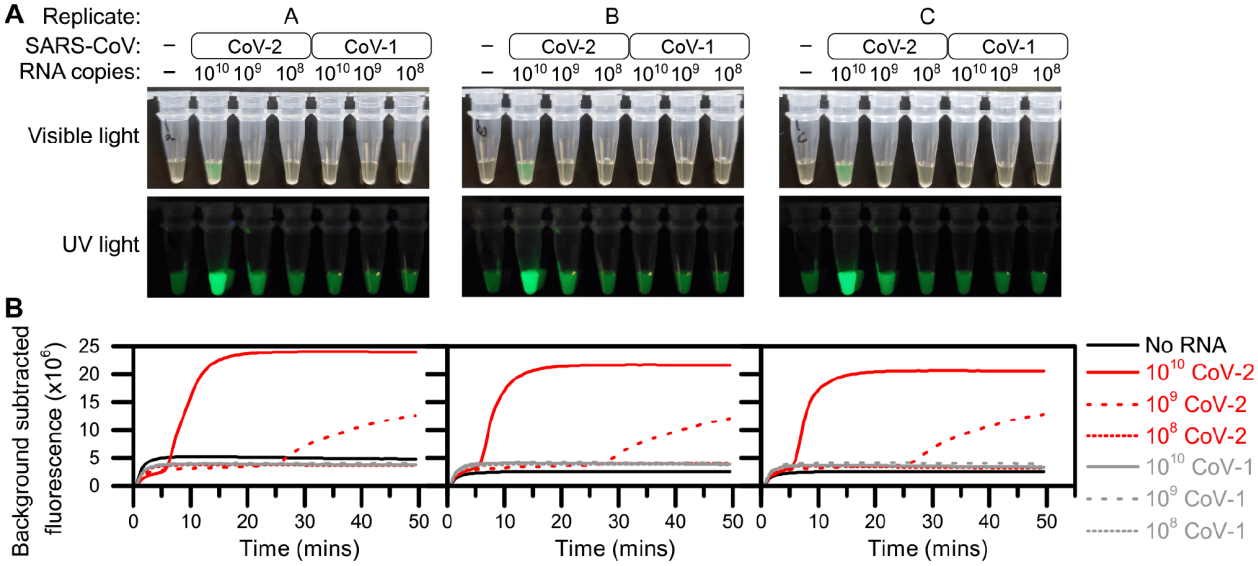


Figure S8.4. Visible fluorometric detection of a specific RNA sequence. Related to Figure 8.1. (A) TtCsm$^{Csm3\text{-}D34A}$ was incubated with no RNA, a specific (i.e., SARS-CoV-2) or non-specific (i.e., SARS-CoV-1) RNA target, in the presence of ATP for 50 minutes at 60ºC. The reactions contain an excess of Mg^{2+} and Mn^{2+} ions, and the metal indicator Calcein. Calcein preferentially binds Mn^{2+} forming a quenched complex. Specific RNA recognition, and associated Cas10-mediated ATP polymerization, generates pyrophosphate which precipitates with Mg^{2+} and Mn^{2+} ions. Excess Mg^{2+} ions bind Calcein to form a highly fluorescent complex that can be seen by eye in visible light or with UV light. (B) Kinetics of Calcein fluorescence for corresponding reactions shown in Panel A. 10^{10} SARS-CoV-2 RNAs can be detected in 10 minutes, and 10^{9} copies are detected in 40 minutes. No increase in fluorescence is seen beyond the first five minutes for either a no RNA control, or for SARS-CoV-21 RNA containing samples.

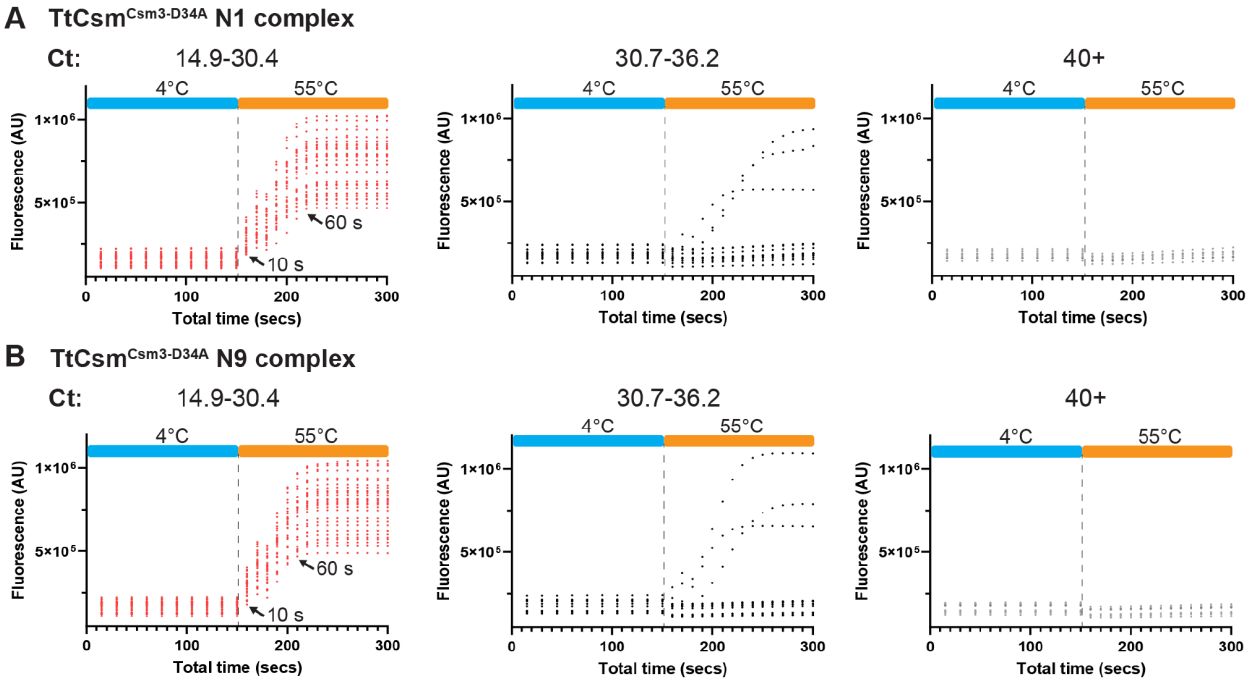


Figure S8.5. Rapid and specific detection of SARS-CoV-2 by RT-LAMP-T7-Csm. Related to Figure 8.3. Raw fluorescence kinetics for the T7-Csm stage of RT-LAMP-T7-Csm detection of SARS-CoV-2 RNA from patient samples with Ct values of 14.9-30.4 (left), 30.7-36.2 (middle), and 40+ (right), as detected by the RNase dead TtCsm complex (dTtCsm) loaded with (A) $crRNA_{N1}$ or (B) $crRNA_{N9}$. Reactions were first incubated in an RT-PCR machine at 4°C and fluorescence was measured every 15 seconds for 150 seconds. Fluorescence readings in both positive and negative reactions are low until the T7-Csm reactions are heated to 55°C, upon which fluorescence rapidly increases within 10 seconds and continues to increase for 60 seconds in most samples.

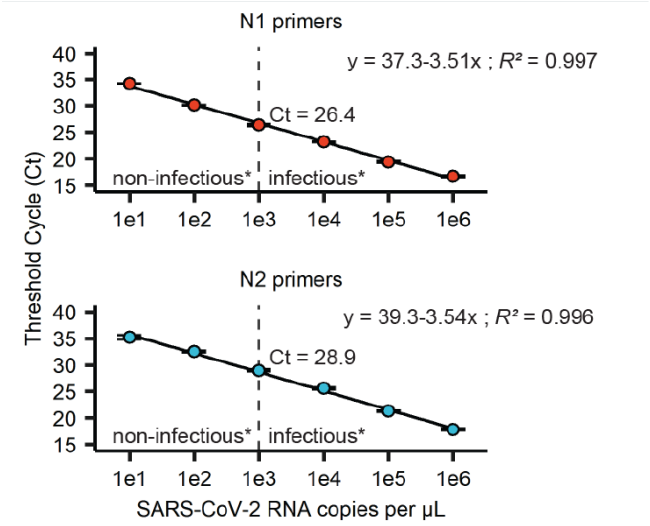


Figure S8.6. Standard curves for absolute quantification of SARS-CoV-2 titers. Related to Figures 8.2 and 8.3. A 10-fold dilution series of SARS-CoV-2 synthetic RNA fragment (RTGM 10169, National Institute of Standards and Technology) containing the Nucleocapsid gene was used. Data was plotted as Cycle Threshold (Ct) on y-axis versus log10 (copies per ml) on x-axis. The mean Ct of three technical replicates are shown, error bars represent ±1 standard deviation. Trend lines were fit to the data using the geom_smooth function of theggplot2 Rpackage; lineare quations and R2 values are shown. Boundaries corresponding to non-infectious and infectious Ct values refer to observations that patients with viral titers below 10^6 copies/ml are rarely infectious.

Supplementary Tables

Name	Sequence (5'-3') [Target binding region underlined]
N1 (76 nt guide)	AUUGCGACACGCUGAAGCGCUGGGGGCAAAUUGUGCAAUUUGCGGCCAGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N3 (76 nt guide)	AUUGCGACGGCCGACGUUGUUUUGAUCGCGCCCCACUGCGUUCUCCAUGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N6 (76 nt guide)	AUUGCGACGUUGCGACUACGUGAUGAGGAACGAGAAGAGGCUUGACUGGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N7 (76 nt guide)	AUUGCGACAGCAGCAGCAAAGCAAGAGCAGCAUCACCGCCAUUGCCAGGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N8 (76 nt guide)	AUUGCGACAUGCUUUAGUGGCAGUACGUUUUUGCCGAGGCUUCUUAGAGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N9 (76 nt guide)	AUUGCGACUUCCGAAGAACGCUGAAGCGCUGGGGGCAAAUUGUGCAAUGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N10 (76 nt guide)	AUUGCGACAUUCAGCAAAAUGACUUGAUCUUUGAAAUUUGGAUCUUUGGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N11 (76 nt guide)	AUUGCGACCAGUUUGCUGUUUCUUCUGUCUCUGCGGUAAGGCUUGAGUGUUGCAAGGGAUUGAGCCCCGUAAGGGG
N12 (76 nt guide)	AUUGCGACGUCAGCACUGCUCAUGGAUUGUUGCAAUUGUUUGGAGAAAGUUGCAAGGGAUUGAGCCCCGUAAGGGG
I1 (76 nt guide)	AUUGCGACAAAAGCGAAAACGUUUAUAUAGCCCAUCUGCCUUGUGUGGGUUGCAAGGGAUUGAGCCCCGUAAGGGG

Table S8.1. TtCsm guide RNAs. Related to Figures 8.1-8.3 and S8.1-S8.5.

Name	Sequence (5'-3')	Associated Figure
RNA Reporter A	/56-FAM/rCrUrCrUrCrU/3IABkFQ/	1
RNA Reporter B	/56-FAM/rArUrCrUrUrCrUrUrArU/3IABkFQ/	2 and 3

Table S8.2. Fluorescent reporter RNAs. Related to Figures 8.1-8.3 and S8.5.

Primer name	Sequence (5'-3')
SARS-CoV-2 N1 T7 template Forward	GATAATACGACTCACTATAGGGAACTGATTACAAACATTG GCCGCAAATTGCACAATT
SARS-CoV-2 N1 T7 template Reverse	GCGCGACATTCCGAAGAACGCTGAAGCGCTGGGGGCAAA TTGTGCAATTTGCGGCC
SARS-CoV-1 N1 T7 template Forward	GATAATACGACTCACTATAGGGAACTGATTACAAACATTG GCCGCAAATTGCACAATT
SARS-CoV-1 N1 T7 template Reverse	GCGTGACATTCCAAAGAATGCAGAGGCACTTGGAGCAAA TTGTGCAATTTGCGGCC
SARS-CoV-2 target all ten complexes_F	CTAGAGCTCGATAATACGACTCACTATAGGGCGTGTTGTTT TAGATTTCATCTAAACG
SARS-CoV-2 target all ten complexes_R	ATCCTGCAGGCACACTGATTAAAGATTGCTATGTG

Table S8.3. Primers to generate IVT RNA templates. Related to Figures 8.1, 8.2, S8.2, S8.3, and S8.4

Primer name	Sequence (5'-3')
F3	GCTGCTGAGGCTTCTAAG
B3	GCGTCAATATGCTTATTCAGC
BIP	TCAGCGTTCTTCGGAATGTCGCTGTGTAGGTCAACCACG
FIP	GCGGCCAATGTTTGTAATCAGTAGACGTGGTCCAGAACAA
Loop Forward	CCTTGTCTGATTAGTTCCTGGT
Loop Reverse	TGGCATGGAAGTCACACC
T7-FIP	TAATACGACTCACTATAGGGAGACGTGGTCCAGAACAA

Table S8.4. Primers for RT-LAMP. Related to Figures 8.3 and S8.5.

Name	Source
SARS-CoV-2	The National Institute of Standards and Technology(RGTM 10169)
SARS-CoV-1	American Type Culture Collection (ATCC) (VR-3280SD)
MERS-CoV	ATCC (VR-3248SD)
Human coronavirus HKU1	ATCC (VR-3262SD)
Influenza B	ATCC (VR-1885DQ)
Human coronavirus NL63	ATCC (VR-3263SD)
Human respiratory syncytial	ATCC (VR-1580DQ)
Pseudomonas aeruginosa	ATCC (27853D-5)
Candida albicans	ATCC (10231D-5)

Table S8.5. Purified genomic nucleic acids. Related to Figure 8.3

Materials and Methods

Bacterial Strains

Escherichia coli DH5a (Thermo Fisher Scientific) cells were used to amplify plasmids used in this paper. *E. coli* BL21 DE3 (NEB) cells were used to express proteins used in this paper. *E. coli* were grown in LB (Lennox) media at either 37ºC, or 16ºC after induction of protein expression with 0.5 mM IPTG (isopropyl-b-D-thiogalactoside).

Nucleic Acid Preparation

Previously published LAMP primers (Eurofins) were designed to amplify the SARS-CoV-2 N-gene.[13] Target SARS-CoV-2 and SARS-CoV-1 RNAs were *in vitro* transcribed with MEGAscript T7 (Thermo Fisher Scientific) from PCR products generated from pairs of synthesized overlapping DNA oligos or using SARS-CoV-2 genome as a template (Table S8.3) (Eurofins). Previously designed primer pools (IDT) were used for RT-PCR and sequencing of SARS-CoV-2 genomes (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop. html). Transcribed RNAs were purified by denaturing PAGE. Fluorescent reporter RNA A and fluorescent reporter RNA B purified by RNase-free HPLC (Table S8.2) (IDT). Purified genomes of viral, bacterial and fungal pathogens were used as is, or resuspended in 1x TE (10 mM Tris-HCl pH 7.5, 1 mM Ethylenediaminetetraacetic acid (EDTA)) to ~1x106 genomes/mL (Table S8.5).

Plasmids

Expression vectors for *Thermus thermophilus* type III-A *csm1-csm5* genes, pCDF-5xT7-TtCsm were purchased from Addgene (plasmid # 128572).[45] pCDF-5xT7-TtCsm was used as a

template for site-directed mutagenesis to mutate the Csm3 residue D33 to alanine (D33A) to inactivate Csm3-mediated cleavage of target RNA (pCDF-5xT7-TtCsmCsm3-D34A).[29] The CRISPR array in pACYC-TtCas6-4xcrRNA4.5 (Addgene plasmid # 127764)[45] was replaced with a synthetic CRISPR array (GeneArt) containing five repeats and four identical spacers, designed to target the N-gene of SARS-CoV-2 (i.e., pACYC-TtCas6-4xgCoV2N1). TtCas6 was PCR amplified from the pACYC-TtCas6-4xcrRNA4.5 plasmid and cloned between the NcoI and XhoI sites of pRSF-1b (MilliporeSigma) (pRSF-TtCas6). The CARF-HEPN nuclease TtCsm6 was expressed from pC0075 TtCsm6 His6-TwinStrep-SUMO-BsaI (Addgene plasmid # 115270).[14]

Protein Purifications

Expression and purification of the TtCsm complex was performed as previously described with minor modifications.[45] Briefly, the crRNA plasmid (e.g., pACYC-TtCas6-4xgCoV2N1) was co-transformed with pRSF-TtCas6 and either pCDF-5xT7-TtCsm or pCDF-5xT7-TtCsmCsm3-D34A into *Escherichia coli* BL21(DE3) cells and grown in LB Broth (Lennox) (Thermo Fisher Scientific) at 37ºC to an OD_{600} of 0.5. Cultures were then induced with 0.5 mM IPTG (isopropyl-b-D-thiogalactoside) for expression overnight at 25ºC. Cells were pelleted (3,000 x *g* for 25 mins at 4ºC) and lysed via sonication in Lysis buffer (25 mM HEPES pH 7.5, 150 mM KCl, 10 mM imidazole, 1 mM TCEP, 0.01 % Triton X-100, 5 % glycerol, 1 mM PMSF). Lysate was clarified by centrifugation at 10,000 x *g* for 25 mins at 4ºC. The lysate was then heat-treated at 55ºC for 45 minutes and further clarified by centrifugation at 10,000 x *g* for 25 mins at 4ºC. His-tagged Csm1 and TtCsm complex were bound to HisTrap HP resin (Cytiva) and washed with Wash buffer (50 mM HEPES pH 7.5, 150 mM KCl, 1 mM TCEP, 5 %

glycerol, 20 mM imidazole). Protein was eluted in Lysis buffer supplemented with 300 mM imidazole. Eluted protein was concentrated (100k MWCO Corning Spin-X concentrators) at 4ºC before further purification over HiLoad Superdex 200 26/600 or Superose 6 Increase 10/300 GL size-exclusion columns (Cytiva) in 25 mM HEPES pH 7.5, 150 mM NaCl, 5% glycerol, 1 mM TCEP. Fractions containing the TtCsm complex were pooled, concentrated, aliquoted, flash frozen in liquid nitrogen, and stored at -80ºC.

Expression and purification of TtCsm6 was performed as previously described with minor modifications.[14] pTtCsm6 was transformed into *Escherichia coli* BL21(DE3) cells and grown in LB Broth (Lennox) (Thermo Fisher Scientific) at 37ºC to an OD_{600} of 0.5. Cultures were then incubated on ice for 1 hour, and then induced with 0.5 mM IPTG for expression overnight at 16ºC. Cells were lysed via sonication in TtCsm6 Lysis buffer (20 mM Tris-HCl pH 8, 500 mM NaCl, 1 mM TCEP) and lysate was clarified by centrifugation at 10,000 x *g* for 25 mins at 4ºC. The lysate was heat-treated at 55ºC for 45 minutes and clarified by centrifugation at 10,000 x *g* for 25 mins at 4ºC. His6-TwinStrep-tagged TtCsm6 was bound to StrepTrap HP resin (Cytiva) and washed in TtCsm6 Lysis buffer. The protein was eluted with TtCsm6 Lysis buffer supplemented with 2.5 mM desthiobiotin and concentrated (10k MWCO Corning Spin-X concentrators) at 4ºC. Affinity tags were removed from TtCsm6 using SUMO protease (100 mL of 2.5 mg/ ml protease per 20 mg of TtCsm6 substrate) during dialysis against SUMO digest buffer (30 mM Tris-HCl pH 8, 500 mM NaCl 1 mM DTT, 0.15% Igepal) at 4ºC overnight. Cleaved His6-TwinStrep tag and uncleaved His6-TwinStrep-TtCsm6 were removed by binding to HisTrap HP resin (Cytiva), and the flow-through was concentrated using Corning Spin-X concentrators at 4ºC. Finally, TtCsm6 was purified using a HiLoad Superdex 200 26/600 size-

exclusion column (Cytiva) in 20 mM Tris-HCl pH 7.5, 1 mM DTT, 400 mM monopotassium glutamate, 5 % glycerol. Fractions containing TtCsm6 were pooled, concentrated, aliquoted, flash frozen in liquid nitrogen, and stored at 80ºC.

To screen guide RNAs in a high throughput format, ten TtCsm complexes were first crudely purified. 8 mL cultures of *E. coli* BL21DE3 cells transformed with pTtCsm and pT7-5xCRISPR-Cas6 were grown at 37ºC and 250 RPM in LB media with selective antibiotics until they reached an OD600 reading of 0.4. Protein expression was then induced with the addition of 0.5 mM IPTG to the media, and cells were grown overnight at 16C. Cells were collected by centrifugation at 4000 RPM, and cell pellets were resuspended in 250 mL of Ni-NTA Equilibration buffer (PBS; 100mM sodium phosphate, 600mM sodium chloride), 0.05% Tween-20 Detergent, 30mM imidazole; pH 8.0). Resuspended cells were sonicated twice for twenty seconds, then clarified by centrifugation at 15,000 rpm for 20 minutes at 4ºC to remove cellular debris. The lysate was then heat-treated at 55ºC for 45 minutes, and re-clarified by centrifugation at 15,000 rpm, for 30 mins at 4ºC. TtCsm was then purified using HisPur Ni-NTA magnetic beads (Thermo Fisher Scientific) according to the manufacturers recommendations, but with modified wash (25 mM HEPES pH 7.5, 150 mM NaCl, 0.05% Tween-20, 1 mM TCEP) and equilibration (25 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP) and elution buffers (25 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP, 300 mM Imidazole). TtCsm complex concentration was quantified on a Nanodrop (Thermo Fisher Scientific).

Type III CRISPR-Csm Based Detection

For experiments shown in Figure 8.1C, RNA was extracted from nasopharyngeal swabs derived from patients that tested negative for SARS-CoV-2 as determined by RT-qPCR. This

RNA was used as is or spiked with *in vitro* transcribed SARS-CoV-2 or SARS-CoV-1 RNA. These RNA samples were mixed with 250 mM ATP, 500 nM fluorescent reporter RNA A, 500 nM TtCsm complex, and 2500 nM of TtCsm6 in reaction buffer (20 mM Tris-HCl pH 7.9, 200 mM Monopotassium glutamate, 10 mM Ammonium sulfate, 5 mM Magnesium sulfate and 1 mM TCEP (tris(2-carboxyethyl)phosphine)) in a 30 mL reaction. Reactions were incubated at 60C (CRISPR-Csm alone), and fluorescence was measured over time in an ABI 7500 Fast Real-Time PCR System (Applied Biosystems), using the manufacturers default filter settings for FAM dye. Fluorescence measurements at an incubation time of 45 minutes are reported.

For experiments shown in Figure 8.2C, an RNA detection mixture was made containing either 25 nM TtCsmCsm3-D34A N1, or 25 nM TtCsmCsm3-D34A N9, or 2.5 nM each of ten complexes (TtCsmCsm3-D34A N1, N3, N6, N7, N8, N9, N10, N11, N12 and I1), mixed with 250 mM ATP, 150 nM fluorescent reporter RNA B, 300 nM TtCsm6, in reaction buffer (20 mM Tris-HCl pH 7.8, 250 mM Monopotassium glutamate, 10 mM Ammonium sulfate, 5 mM Magnesium sulfate and 1 mM TCEP). 3 mL of RNA extracted from a patient nasopharyngeal swab with high SARS-CoV-2 viral load (~5x108 copies/mL) was added to 27 mL of the above RNA detection mixture. Alternatively, RNA from this positive patient sample was first diluted 10– or 100–fold into RNA extracted from a patient negative for SARS-CoV-2 (CT > 40), and 3 mL of these dilutions was added to 27 mL of the above RNA detection mixture. Reactions were incubated at 60ºC and fluorescence was measured every 10 s for up to 20 minutes, in a QuantStudio 3 Real-Time PCR system (ThermoFisher), using the manufacturers default filter settings for FAM dye.

Colorimetric CRISPR-Csm based detection

TtCsm$^{Csm3\text{-}D34A}$ stocks were buffer exchanged into a low buffering capacity buffer (0.5 mM Tris-HCl pH 8.8, 50 mM Potassium chloride, 10 mM Ammonium sulfate, 8 mM Magnesium sulfate) using Microspin G25 columns (Cytiva) as per the manufacturer's instructions. TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA) or *in vitro* transcribed SARS-CoV-2 or SARS-CoV-1 RNA were incubated with 200 nM TtCsm$^{Csm3\text{-}D34A}$ in 1x WarmStart Colorimetric LAMP Master Mix (NEB), supplemented with an additional 1 mM ATP, in a 25 mL reaction. The volume of buffer-exchanged TtCsm used contributed approximately 40 mM Tris-HCl pH 8.8 buffer to the final reaction. Reactions were assembled on ice and imaged on an LED tracing pad with a Galaxy S9 phone (Samsung). Then reactions were incubated at 60C for 30 minutes, rapidly cooled, and imaged again.

Visible Fluorometric CRISPR-Csm Based Detection

TE buffer or *in vitro* transcribed SARS-CoV-2 or SARS-CoV-1 RNA were incubated with 500 nM TtCsm$^{Csm3\text{-}D34A}$ in reaction buffer (20 mM Tris-HCl pH 8.8, 100 mM Potassium chloride, 10 mM Ammonium sulfate, 6 mM Magnesium sulfate, 0.5 mM Manganese chloride, 1 mM TCEP, 1 mM ATP and 25 mM Calcein), in a 30 mL reaction. Reactions were incubated at 60C, and fluorescence was measured over time in an ABI 7500 Fast Real-Time PCR System (Applied Biosystems), using the manufacturers default filter settings for FAM dye. After incubating at 60ºC for 50 minutes, the same reactions were then imaged under visible light, and under UV light (365 nm) with a Galaxy S9 phone (Samsung). To screen guide RNAs in a high throughput format (Figure 8.2B), 200 nM crude purified TtCsm complex was incubated with 10^{12} copies of IVT SARS-CoV-2 RNA in the above buffers and fluorescence was recorded in a ABI 7500 Fast Real-Time PCR System (Applied Biosystems) machine as above.

RT-LAMP-T7-Csm

Isothermal amplification of nucleic acids in swab samples was performed by RT-LAMP. In brief, 25 mL reactions contained 8 units (U) of WarmStart Bst 2.0 (NEB), and 7.5 U of WarmStart RTx Reverse Transcriptase (NEB), 1.4 mM dNTPs, LAMP primers, 25 U of Murine RNase Inhibitor (NEB) in reaction buffer (20 mM Tris-HCl pH 7.8, 8 mM Magnesium sulfate, 10 mM Ammonium sulfate, 50 mM potassium chloride, 0.1% Tween-20). LAMP primers designed to amplify the SARS-CoV-2 N-gene,[13] were added at an optimized final concentration of 0.2 mM F3 and B3, 0.4 mM LoopF and LoopB, 1.6 mM BIP, 0.53 mM FIP, and 1.07 mM of T7-FIP (Table S8.4). The T7-FIP primer consists of a T7 promoter fused to the 5' end of the FIP primer, and allows for the generation of T7 transcription templates during the second step of T7-Csm reaction. RT-LAMP reactions were performed using 5 mL of input RNA at 65ºC for 29 minutes. 3 mL of RT-LAMP reactions were mixed with 27 mL of a modified T7-Csm fluorescent detection reaction containing 0.5 mM rNTPs, 300 nM TtCsm6, 5.5 units of Hi-T7 RNA Polymerase (NEB), 150 nM fluorescent reporter RNA B, and 20 nM of either $TtCsm^{Csm3\text{-}D34A}$ N1 or N9, in reaction buffer (40 mM Tris-HCl pH 7.5, 4 mM Magnesium chloride, 50 mM Sodium chloride, 2 mM spermidine, 1 mM DTT). Reactions were incubated at 55ºC for up to 20 min and fluorescence kinetics was monitored in a QuantStudio 3 Real-Time PCR system (ThermoFisher) as described above.

LoD standards were prepared by diluting SARS-CoV-2 RNA into RNA extracted from COVID-19-negative patient nasopharyngeal swabs. Concentrations were determined with RT-qPCR using a standard curve generated from 10-fold dilution series ($1x10^6$-$1x10^0$) of IVT fragment.

Human Clinical Sample Collection and Preparation

Nasopharyngeal swabs from patients that either tested negative or positive for SARS-CoV-2 were collected in viral transport media. RNA was extracted from all patient samples using QIAamp Viral RNA Mini Kit (QIAGEN).

RT-qPCR

RT-qPCR was performed using two primers pairs (N1 and N2) and probes from the 2019-nCoV CDC EUA Kit (IDT#10006606). SARS-CoV-2 in RNA-extracted, nasopharyngeal patient samples was detected and quantified using one-step RT-qPCR in ABI 7500 Fast Real-Time PCR System according to CDC guidelines and protocols (https://www.fda.gov/media/134922/download). In brief, 20 mL reactions included 8.5 mL of Nuclease-free Water, 1.5 mL of Primer and Probe mix (IDT, 10006713), 5 mL of TaqPath 1-Step RT-qPCR Master Mix (ThermoFisher, A15299) and 5 mL of the template. Nuclease-free water was used as negative template control (NTC). Amplification was performed as follows: 25ºC for 2 min, 50ºC for 15 min, 95ºC for 2 min followed by 45 cycles of 95ºC for 3 s and 55ºC for 30 s. To quantify viral genome copy numbers in the samples, standard curves for N1 and N2 were generated using a dilution series of a SARS-CoV-2 synthetic RNA fragment (RTGM 10169, National Institute of Standards and Technology) spanning N gene with concentrations ranging from 10 to 10^6 copies per mL. Three technical replicates were performed at each dilution. The NTC showed no amplification throughout the 45 cycles of qPCR.

Bioinformatic Design of TtCsm crRNA Guides Targeting SARS-CoV-2

An alignment of 45,641 SARS-CoV-2 genomes was downloaded from the GISAID database (Global Initiative for Sharing All Influenza Data; https://www.gisaid.org/ on 6-23-

2020.[34,46] The alignment was scanned for conservation with a 40-nucleotide sliding window, and 40-nucleotide segments with strong conservation were saved for downstream analysis. Next, four nucleotides flanking the above 40-nucleotide candidate viral target sequences were checked for base pairing to the first four nucleotides of the prospective 5'-crRNA handle (underlined; 5'-AUUGCGAC-3'), only candidates lacking handle complementarity were considered further. Candidate sites with less than two mismatches to SARS-CoV (NC_004718.3) and MERS-CoV (NC_019843.3) in the first 18 nucleotides of the target sequence were discarded. Next, candidate crRNAs targeting the above sites were screened for potential cross-reactivity with human mRNAs and a list of human pathogens and common respiratory flora downloaded from the FDA's Emergency Use Authorization requirements (downloaded on 7-29-2020) using BLAST (E-value 1000). The remaining 6,229 crRNA sequences were then sorted by genomic location and only guides that were located 3^0 of the SARS-CoV-2 ORF3a gene (positions 25,393 to 29,903) were considered further. Finally, 76 guides were selected from the remaining pool that had the greatest conservation among SARS-CoV-2 sequences and the largest number of mismatches to SARS-CoV and MERS-CoV sequences.

Sequencing of SARS-CoV-2 RNA Isolated from Patient Samples

SARS-CoV-2 genomic RNA isolated from patient samples was sequenced as previously described.[47] In brief, 10 mL of SARS-CoV-2 genomic RNA extracted from nasopharyngeal patient swabs was first reverse transcribed with SuperScript IV (ThermoFisher) according to the manufacturer's instructions. The ARTIC Network protocol was followed to generate a sequence amplicon library covering the whole SARS-CoV-2 genome on Oxford Nanopore using a ligation sequencing kit (Oxford Nanopore, SQK-LSK109) (https://artic.network/ncov-2019/ncov2019-

bioinformatics-sop.html).[48,49] Two multiplex PCR reactions were performed with primer pools described in the ARTIC nCoV-2019 V3 Panel, amplified with Q5 DNA Polymerase (NEB). The two resulting amplicon pools for each patient sample were then combined and used for library preparation. Samples were end repaired (NEB, E7546) and then barcoded using Native Barcoding Expansion Kits (Oxford Nanopore, EXP-NBD104 and EXP-NBD114). Barcoded samples were pooled together and then Nanopore adaptors were ligated.

The multiplexed library was loaded onto the MinION flowcell, and a total of 0.3 Gb of raw sequencing data was collected per patient sample. Raw Nanopore reads were base-called in high-accuracy mode (Oxford Nanopore, MinKNOW), and further analyzed using the ARTIC bioinformatic pipeline for COVID-19 (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html).[50] Consensus sequences were uploaded to GISAID (https://www.gisaid.org/), IDs: EPI_ISL_1081321, EPI-ISL_1081322, EPI_ISL_1081323.[34] These three SARS-CoV-2 genome sequences were identified as members of the B.1.1.7 lineage by an automated lineage assigner (https://github.com/cov-lineages/pangolin).[51]

Quantification and Statistical Analysis

All experiments were performed in triplicate or duplicate and error is reported as ± 1 standard deviation. The merged datasets of replicates of fluorescence kinetics of direct Csm-based detection of SARS-CoV-2 RNA in patient samples was fit to a simple linear regression, in Prism 9 (Graphpad). The fitted slopes of SARS-CoV-2 RNA-containing patient samples were compared pairwise to the negative swab RNA control by an F-test, ****$p < 0.0001$.

CHAPTER NINE

FUNCTIONAL AND PHYLOGENETIC DIVERSITY OF CAS10 PROTEINS

Contribution of Authors and Co-Authors

Manuscript in Chapter 9

Author: Tanner Wiegand

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript, wrote the manuscript

Co-Author: Royce Wilkonson

Contributions: purified proteins

Co-Author: Andrew Santiago-Frangos

Contributions: performed the research

Co-Author: Mackenzie Lynes

Contributions: acquired samples and performed metagenomic sequencing

Co-Author: Roland Hatzenpichler

Contributions: acquired samples and performed metagenomic sequencing

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

Tanner Wiegand, Royce Wilkonson, Andrew Santiago-Frangos, Mackenzie Lynes, Roland Hatzenpichler, Blake Wiedenheft

The CRISPR Journal

Status of Manuscript: [Put an x in one of the options below, delete this]
__X_ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
____ Published in a peer-reviewed journal

FUNCTIONAL AND PHYLOGENETIC DIVERSITY OF CAS10 PROTEINS

Tanner Wiegand,[1] Royce Wilkonson,[1] Andrew Santiago-Frangos,[1] Mackenzie Lynes,[2] Roland Hatzenpichler,[2] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717, USA.

Abstract

Cas10 proteins are large subunits of type III and I-D CRISPR-RNA (crRNA) guided surveillance complexes, many of which have nuclease and polymerase activities. Here, we identify and analyze 8,088 Cas10 sequences from genomic and metagenomic databases. These analyses reveal the evolutionary relatedness of Cas10 proteins and show that only a quarter of them have conserved HD nuclease active site motifs, while active site motifs in Cas10 polymerase domains are more widely conserved. We identify Cas10 variants that are fused to nucleases that are activated by cyclic nucleotides (i.e., NucC from cBASS systems) and components of toxin-antitoxin systems (i.e., AbiEii from AbiE systems), showing how CRISPR-Cas systems become diversified and gain new putative mechanisms of immunity. Finally, we clone five Cas10 representatives from three phylogenetically distinct clades. We purify the proteins and show that all of them have DNA-dependent DNA polymerase activity. Collectively, this work reveals the phylogenetic diversity of Cas10 proteins and demonstrates the functional parallels of five distantly related Cas10 homologs.

Introduction

Type III CRISPR-Cas systems are divided into six phylogenetically distinct subtypes (i.e., A-F), most of which rely on four to six Cas proteins that assemble around a processed CRISPR RNA guide (crRNA) to form a large multi-subunit surveillance complex.[1-6] These crRNA-guided surveillance complexes, are called Csm (subtypes III-A and III-D) or Cmr (subtypes III-B and III-C) proteins due to historical naming conventions. Cas10 proteins (i.e., Csm1 or Cmr2) are large subunits of Csm/Cmr complexes that until recently—with the recognition of Cas10-less type III-E systems—were signature proteins universal to all type III systems.[5,7]

Type III Cas10 proteins are comprised of an N-terminal HD nuclease domain, two Palm domains that are separated by a zing-finger motif, and a C-terminal domain that anchors Cas10 to the Csm/Cmr complex (Figure 9.1A).[2,3,6,9] Single-stranded DNase activity has been reported for the HD domains of some Cas10 proteins, which is thought to promote nicking of ssDNA formed in transcription bubbles.[10-12] While the HD-mediated ssDNA nicking activity of Cas10 has been shown to promote mutagenesis of Staphylococci, allowing rapid responses to environmental perturbations (e. g., antibiotic selection), Cas10 HD activity has been shown to be dispensable for immunity in some type III systems.[13,14] Indeed, numerous Cas10 proteins lack an N-terminal HD domain, although the abundance and phylogenetic distribution of these truncated proteins has not been reported.[4,5]

In 2017, two groups independently showed that target RNA recognition activates the Cas10 cyclase, which converts ribo-ATP into cyclic oligoadenylates (cOA).[15,16] These Cas10-produced cOA-messengers go on to activate trans-acting effectors (e.g. Csm6 or Card1) that

indiscriminately cleave RNA or DNA, stalling replication of both the host and phage.[15,17] Cyclization of ATP relies on a di-glycine di-aspartate (GGDD) motif, that is conserved in one of the two Palm domains of Cas10.[18,19] In addition to cOA-synthesis, recently published data indicate that some Cas10 proteins may also have DNA-dependent DNA polymerase activity.[20] Degeneration of GGDD motifs has been reported for some type III-associated Cas10 proteins (i.e., some III-C and III-F), but like HD motifs, it remains unclear how widespread GGDD motifs are in type III Cas10 proteins.[4,5]

Here, we show that Cas10 proteins from the same CRISPR subtype form monophyletic clades in a phylogenetic tree composed of 8,088 Cas10 sequences. This analysis reveals a more ancient origin for III-B Cas10 proteins and indicates that Cas10 sequences from III-A and III-D systems arose most recently. These data also reveal that Cas10 proteins are occasionally fused with components of other immune systems (e.g., NucC from cBASS) and toxin-antitoxin systems (e.g., AbiEii from AbiE). To better understand the functional diversity of Cas10 proteins, we analyze each Cas10 sequence and show that HD nuclease active site motifs are only conserved in about a quarter of Cas10 proteins, while Palm2 GGDD motifs are strongly conserved in Cas10 proteins from every type III subtype except III-C and III-F. Intrigued by the widespread conservation of Cas10 GGDD motifs—and the recent report of Cas10 DNA-dependent DNA polymerase activity[20]—we test five phylogenetically distant Cas10 proteins *in vitro* and show that each protein has DNA-dependent DNA polymerase activity.

Materials and Methods

Identification of Cas10 proteins

Prodigal was used to translate all open reading frames (ORFs) in 58,864 genomes downloaded from the NCBI database, 9,687 genomes from the Joint Genomics Institute (JGI), and 26,354,826 metagenomic scaffolds from samples isolated in Yellowstone National Park. Profile hidden markov models (HMMs) of Cas10 were pulled from CasFinder and used to query the entire database.[22] To identify III-F Cas10 proteins, we also queried the database with an additional profile HMM that was constructed from eleven III-F Cas10 sequences identified by Makarova *et al*.[5] HMM queries were conducted in hmmer with an E-value cutoff of 0.01.[23] Search results were further filtered in R to exclude matches that did not meet a sequence E-value threshold of 1e-05. This search resulted in the identification of 10,128 Cas10-like sequences. Redundant sequences were filtered using CD-HIT, resulting in a set of 8,088 unique Cas10-like proteins.[24]

Phylogenetic and domain composition analysis

8,088 unique Cas10-like sequences were aligned in MAFFT using the "auto" setting, and the resulting alignment was used to construct a phylogenetic tree in FastTree with -wag and -gamma model settings.[25,26] Phylogenetic trees were visualized using the ggTree package from BioConductor in R, after being rooted using the 'midpoint' function in the APE package.[27,28] Each HMM used for initial Cas10 identification was composed of subtype specific Cas10 protein sequences (e.g., III-A Cas10s only), which enabled subtypes to be assigned to each Cas10-like sequence. To identify III-A- and III-B-like Cas10 features present in each sequence, we built an HMM of sequences of five structurally determined Cas10 proteins: TonCsm1 (*Thermococcus onnurineus*), SthCsm1 (*Streptococcus thermophilus*), TthCsm1 (*Thermus thermophilus*), SsoCsm1 (*Sulfolobus solfataricus*), PfuCmr2 (*Pyrococcus furiosus*). This HMM was then used

to query each Cas10-like sequence, and HMM coverage was assessed in R. Coordinates for the boundaries of each domain in the HMM were determined by manually cross-referencing structures of TonCsm1 (PDB: 4UW2 and 6O73) with results from an HMM search of the TonCsm1 sequence. To determine if ancillary domains were present in Cas10-like sequences, we used the PFAM database and profile HMMs of Cas proteins extracted from the CasFinder program to annotate each Cas10-like sequence.[22,29] Results were manually analyzed, and only novel Cas10 gene fusions identified by a search result with a sequence E-value less than 1e-10 were reported. Structural models shown in Figure 9.1F were predicted with the AlphaFold 'Predict' utility in ChimeraX.[30]

Determining the presence of HD and Palm2 catalytic motifs

To determine if conserved HD and Palm2 motifs were present in the catalytic sites of the Cas10-like sequences identified, we constructed multiple sequence alignments of Cas10 proteins for each CRISPR subtype (i.e., III-A, -B, -C, -D, -F and I-D) using the DECIPHER package in R.[31] Each alignment also included the sequence of TonCsm1 to enable cross-referencing to the TonCsm1 structure (PDB: 4UW2). Each alignment was manually inspected to determine the equivalent positions of residues composing catalytic motifs, before those columns were extracted from each alignment and analyzed in R.

Cas10 protein purification

Five Cas10 homologs (TonCsm1, SthCsm1, TthCsm1, SsoCsm1, and PfuCmr2) were each cloned into a pRSFDuet-1 expression vector. Cas10 genes were codon optimized for expression in *E. coli* and include N-terminal 6x-histidine, 2x strep-affinity, and SUMO solubilization tags. Using this same approach, an additional vector was also constructed

containing TonCsm1 (with an N-terminal six histidine tag) and TonCsm4, as previously described by Jia et al.[2] Protein expression was induced in one-liter cultures of *E. coli* BL21-DE3 when they reached an OD_{600} reading of 0.4-0.5 via addition of 0.2M IPTG. Protein expression continued overnight at 16ºC before cells were collected via centrifugation at 3000 x g (10 minutes at 4ºC). Cells were lysed in 20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 5% glycerol, and 1 mM DTT via sonication, centrifuged at 10,000 x g (25 minutes at 4ºC), and soluble fractions of each protein were bound to Streptactin resin (IBA Lifesciences). Contaminants were removed with 20-30 column volumes of Wash Buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 5% glycerol, 1 mM DTT, 5 mM ATP, 20 mM MgCl2), and bound proteins were eluted with a linear gradient from Wash Buffer to Elution Buffer (20 mM Tris-HCl, pH 8.0, 250 mM NaCl, 5% glycerol, 1 mM TCEP, 2.5 mM desthiobiotin). Affinity tags were cleaved via treatment with SUMO protease overnight at 4ºC. Affinity tags and uncleaved proteins were removed by passing the reaction through a Ni-affinity column and the flow-through was concentrated (10k MWCO concentrator) before final purification over a Superdex 200 column (20 mM Tris-HCl, pH 8.0, 250 mM NaCl, 5% glycerol, 1 mM DTT).

Cas10 cyclase activity assays

To assess the cyclase (i.e. cOA-synthase) activities of five Cas10 proteins, we assembled reactions comprised of 1µM Cas10, 25µM rATP, 0.08µM P^{32}-α-rATP, and a reaction buffer (20mM Tris-HCL pH 7.8, 250mM K-Glutamate, 10mM Ammonium Sulfate, 5mM Magnesium sulfate, 1mM TCEP). Reactions were incubated at either 37ºC (SthCsm1) or 60ºC (TthCsm1, SsoCsm1, TonCsm1, PfuCmr2) for 1 hour. As positive controls, full TthCsm and SthCsm complexes—with crRNAs targeting the N gene of SARS-CoV-2—were purified using a

previously described protocol.[32] TthCsm and SthCsm reactions were prepared as described above, with the addition of 10^{12} copies of *in vitro* transcribed RNA containing the SARS-CoV-2 N gene sequence. Products were extracted from all reactions via phenol-chloroform precipitation and were diluted 2:1 with 100mM Sodium Acetate pH 5.2. Reaction products were separated using thin-layer chromatography and plates were imaged using a Typhoon V (Cytiva) for phosphor-imaging.

Cyclase activity was additionally assessed by incubating Cas10 proteins in a Calcein-containing buffer (20mM Tris-HCl pH 8.8, 100mM Potassium Chloride, 10mM Ammonium Sulfate, 6mM Magnesium Sulfate, 0.5mM Manganese Chloride, 1mM TCEP, 25μM Calcein) with 1mM dNTP, rNTP or rATP alone. Reactions were held at 4ºC for 150 seconds in a QuantStudio RT-qPCR thermocycler (ThermoFisher), before being heated to 37ºC and held at this temperature for one hour. Fluorescence was measured every ten seconds using a FAM-detection setting. Fluorescence data were quantified and analyzed in R.

Cas10 polymerase activity assays

To evaluate the DNA-dependent DNA polymerase activity of five Cas10 proteins, we first annealed a twenty nucleotide 5'-FAM-labelled DNA primer (5'-ACTGCTAATAACGAGCGTTG-3') to a forty nucleotide DNA template (5'-TATCATTGGCTCCTTCAATCCAACGCTCGTTATTAGCAGT-3') at a ratio of 1:3 (20nt-primer:40nt-template). Reactions composed of 1μM Cas10, 10nM DNA-primer DNA-template, and a buffer (20mM Tris-HCL pH8.0, 100mM NaCl, 2.5mM MgCl2, 10% (v/v) glycerol, 2mM DTT, 250μM dNTP, 0.05μg/mL BSA) were incubated at 37ºC for an hour. As a positive control, primer-template substrates were similarly incubated with Q5 DNA polymerase in a 5x Q5 buffer

(NEB) for 10 minutes at 50ºC. The reactions were stopped by adding loading buffer (30% glycerol and 100mM EDTA). Samples were heat denatured for five-minute at 95ºC and then cooled on ice prior to separation using a 19% PAGE gel containing 7M urea. A reaction using the same conditions but substituting a 5'-Cy5-labelled RNA primer (5'-ACUGCUAAUAACGAGCGUUG-3'), was used to assess the DNA-dependent RNA-polymerase activity of Cas10 proteins. Similarly, to test the RNA-dependent RNA polymerase and reverse transcriptase activities of Cas10 proteins, a forty nucleotide RNA template (5'-UAUCAUUGGCUCCUUCAAUCCAACGCUCGUUAUUAGCAGU-3') was used in the above reaction setup with the 5'-Cy5-labelled RNA primer and the 5'-FAM-labelled DNA primer, respectively. PAGE gels were imaged on a Typhoon V, using Cy2 (FAM) and Cy5 lasers.

Results

Phylogenetic diversity and proposed evolutionary history of Cas10 proteins

To investigate the phylogenetic diversity of Cas10 proteins, we used profile hidden markov models (HMMs) to identify Cas10 proteins in 68,551 genomes from the National Center for Biotechnology (NCBI) and the Joint Genomics Institute (JGI) databases. In addition, we also queried 26,354,826 metagenome-scaffolds from environmental samples isolated in Yellowstone National Park. Collectively, these queries identified 8,088 unique Cas10-like sequences which were aligned and used to construct a phylogenetic tree (Figure 9.1B). This analysis reveals six phylogenetically distinct clades that largely recapitulate the CRISPR-Cas subtype designations

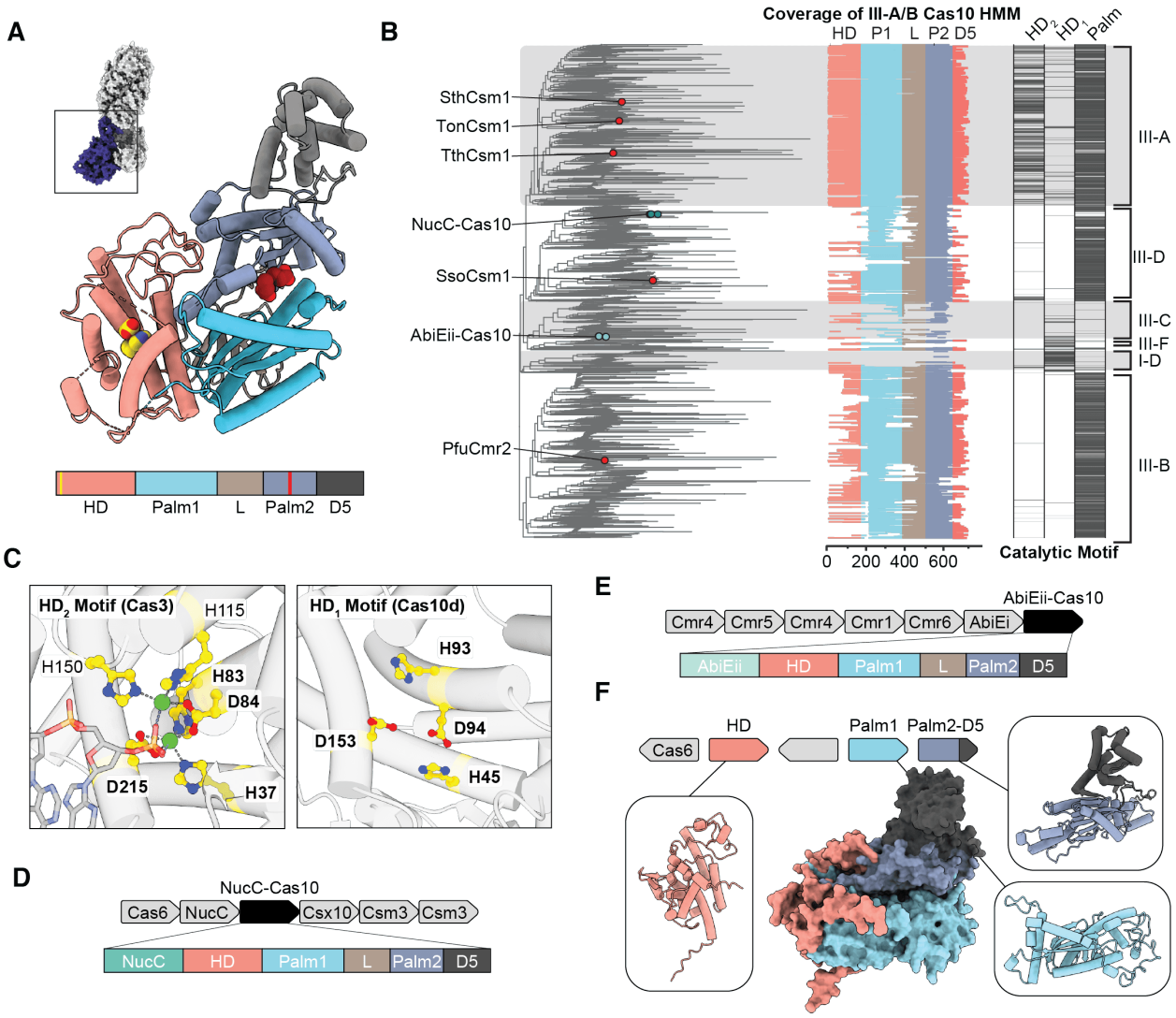


Figure 9.1. Cas10 proteins have variable domain compositions, with subtype-specific patterns. (A) Cas10, shown in blue in upper left inset, is the large subunit of type III and I-D surveillance complexes (PDB: 6O1O). The structure of TonCas10 reveals five domains that are present in most structurally determined Cas10 proteins (PDB: 4UW2). HD and Palm2 active site residues are shown in yellow and red, respectively. (B) Phylogenetic reconstruction of 8,088 Cas10 sequences recapitulates the recently proposed type III and I-D subtype classifications.[5] Coverage of HMM constructed from five structurally determined Cas10 proteins shown as multi-colored columns. Presence of catalytic motifs and subtype classifications shown in right-most columns. (C) Left: Catalytic HDHHDH residues of *T. fusca* Cas3 HD active site (HD_2), shown in yellow, reveal coordination of two metal ions (PDB: 4QQW). Right: Catalytic HHDD residues of *S. islandicus* Cas10d HD active site (HD_1), shown in yellow, coordinate a single metal ion (PDB: 6YES). (D) Newly identified NucC genetically fused to N-terminus of Cas10 shown as schematic protein representation below scheme of encoding locus (JGI genome: 2654587738-2654604082 Ga0100770_105). (E) Newly identified AbiEii genetically fused to N-terminus of Cas10 shown as schematic protein representation below scheme of encoding locus (NCBI

genome: NC_022792.1). (F) Cas10 from *T. marina* is split over three genes, with Cas10 domains defining the boundaries of each gene. AlphaFold predictions of each gene translation were superimposed on TonCsm1 and appear to form a Cas10 complex that is shown as a protein surface representation (NCBI genome: NZ_CP042909.1).

proposed by Koonin's group.[4,5] Cas10 sequences from type III-B systems form a monophyletic clade that emerges from the bifurcated root of the tree, indicating that III-B Cas10s diverged from other Cas10 proteins long ago (Figure 9.1B). In contrast, Cas10 proteins from III-A and III-D systems form monophyletic clades with a shared ancestral node far from the root of the tree, indicating that these proteins emerged from a single ancestor relatively recently. Cas10 proteins from type III-F and III-C systems also form monophyletic clades that emerge from a shared node, suggesting that are closely related. While little is known about immunity in type III-F systems, the close relationship between III-F and III-C Cas10 proteins suggests these systems may have mechanistic similarities.

In addition to Cas10 proteins from type III systems, we also identified Cas10s from type I-D systems (i.e. Cas10d). These proteins contain divergent Palm domains, and previous work suggests that these proteins do not produce cOAs. Instead these proteins rely on Cas10 HD-nuclease activity for clearance of mobile genetic elements.[8,33] I-D Cas10s form a monophyletic clade that is closely related to Cas10s from the III-C and III-F systems (Figure 9.1B). This evolutionary relationship may explain why most III-C Cas10 proteins have mutations in the Palm domain that are anticipated to prevent cOA-synthesis.[4]

Differences in Cas10 HD and Palm active sites

The majority of Type III Cas10 proteins that have been biochemically or biologically investigated to date contain an HD nuclease and cOA-synthesizing Palm domains.[1-3,6,9,15,34] To

assess how widespread these features are in Cas10 proteins, we evaluated Cas10 sequences to determine whether or not each sequence contained HD or Palm active site motifs. HD nuclease domains are found in CRISPR-associated proteins from other types of systems (e.g. Cas3 from type I) and cleave nucleic acid substrates with a histidine- and aspartate-rich active site. Structures of Cas3 proteins from *Thermobifida fusca* and *Methanocaldococcus jannaschii* reveal HD active site coordination of two metal ions via an HDHHDH (i.e., HD_2) motif (Figure 9.1C).[35,36] By contrast, the HD domain of Cas10d from *Sulfolobus islandicus* is reported to coordinate only one metal ion, using an HHDD (i.e., HD_1) motif (Figure 9.1C). Structure-guided analyses of Cas10 multiple sequence alignments reveal that only 25.7% (2,082 of 8,088 sequences) of Cas10 proteins have either an HD_2 (HDHHDH) or HD_1 (HHDD) motif (Figure 9.1B and Table S9.1).[8] This finding implies that HD activity is unlikely to be found in the majority of type III and I-D Cas10 proteins. In agreement with these numbers, 67.6% of Cas10s were found to have truncated or missing HD domains (5,470 of 8,088 sequences) when Cas10 sequences were scanned with profile HMMs built from Cas10 HD domains (Figure 1B).

Cas10 HD domains also differ based on their associated type III subtypes. HD_2 motifs were found in roughly one third (31.0%) of III-A Cas10s, while another 34.8% of III-A Cas10 sequences had HD_1 motifs (Figure 9.1B and Table S9.1). III-A Cas10 proteins with HD_1 motifs were scattered throughout the III-A clade, suggesting that mutations of the active site residues responsible for coordinating a second metal ion have likely arisen several times independently during the evolution of III-A Cas10 proteins. HD_2 motifs were completely absent from all other Cas10 subtypes, indicating that HD coordination of a second metal ion likely emerged in Cas10 proteins from type III-A systems. HD_1 motifs are sparse in III-C Cas10 sequences (present in

7.8%) and were totally absent in III-D Cas10s, suggesting that these systems rely on alternate mechanisms of Cas-mediated immunity (e.g., cOA-gated effectors or Csm3-mediated target RNA cleavage) (Figure 9.1B and Table S9.1).

Two Palm domains are present in most Cas10 proteins, with the C-terminal Palm domain (Palm2) containing the GGDD motif that coordinates divalent cations during cOA synthesis.[2,15,34] Cas10 sequence alignments reveal GGDD motifs in the majority of III-A, III-B, and III-D proteins, consistent with prior analyses of these type III systems (Figure 9.1B and Table S9.1).[4] However, cOA-production is not universal to all Cas10 proteins. One previously determined structure of a Cas10d homolog from *S. islandicus* was shown to have a divergent Palm2 domain, which lacked a GGDD motif and was therefore unable to produce cOA messengers.[8] Mutations in the GGDD motif of the Palm2 domain are a conserved feature of all Cas10d sequences (Figure 9.1B and Table S9.1). GGDD motifs are also absent in III-C and III-F Cas10 sequences, suggesting cOA-synthesis may be a feature that was lost in a shared ancestor of III-C and III-F Cas10s (Figure 9.1B and Table S9.1).

Variations of Cas10 domain architecture

To determine if Cas10 proteins contain additional domains, we annotated each unique Cas10 sequence using profile HMMs of known Cas proteins and the PFAM database.[22,29] This analysis revealed genetic fusions of Cas10 to NucC, a second messenger-gated nuclease. NucC is reported to be a widespread cA_3-activated effector of cyclic oligonucleotide-based anti-phage signaling systems (CBASS).[37] Four sequences were identified with a ~300 residue NucC domain fused to the N-terminal end of Cas10 (i.e., NucC-Cas10) (Figure 9.1D). Unlike most other Cas10s in the III-D clade (Figure 9.1B), the four NucC-Cas10 homologs identified have a

divergent DGDD motif in their Palm2 domains, which is still expected to facilitate conversion of ATP into numerous species of cOA. Other Csm/Cmr complexes have been shown to produce cA_3, suggesting that genetic linkage of NucC to Cas10 in these four newly identified systems may allow funneling of cA_3 products directly from the crRNA-guided surveillance complex to a physically tethered cA_3-gated nuclease effector.[15]

In addition to NucC-Cas10 sequences, our search for novel Cas10 fusions yielded four examples of a type IV toxin from a toxin-antitoxin system (i.e., AbiEii) fused to the N-terminus of Cas10 (Figure 9.1E). AbiEii is a predicted nucleotidyltransferase that specifically binds GTP.[38] While the exact mechanism of AbiEii-mediated toxicity remains unknown, genetic linkage of this protein to Cas10 implies that this toxin may have been coopted for immune purposes. Effector toxicity is a common mechanism utilized by type III CRISPR systems to induce cellular dormancy or death (i.e., abortive infection) during bacteriophage infection.[15,39] We predict that conformational changes induced by AbiEii-Cmr target binding may activate the toxic nucleotidyltransferase activity of AbiEii-Cas10 proteins. Interestingly, the four Cas10s linked to AbiEii (i.e., AbiEii-Cas10) fall into the III-C clade and accordingly do not have conserved GGDD motifs in their Palm2 domains (Figure 9.1B). This suggests that these AbiEii-Cas10s are unlikely to produce cOA molecules, implying that these variant systems do not rely on cOA-gated effectors for immunity.

In several instances, the domains of Cas10 are encoded by separate genes. For example, Cas10 in the III-A system of *Thermosulfurimonas marina* (strain SU872), is split into three genes (Figure 9.1F). Comparison of these three Cas10 sequences with structurally determined Cas10 proteins indicates that each III-A *T. marina* gene starts and ends near the boundary of an

expected Cas10 domain. One gene contains an HD domain, another Palm1, and the third a Palm2 and C-terminal domains (Figure 9.1F). The existence of a standalone Cas10 HD protein suggests a pathway for the evolution of novel Cas domain fusions, supporting the proposed evolutionary path of Cas3 proteins via genetic fusion of a Cas10 HD domain and an SF2 helicase domain.[4] If these Cas10 sub-proteins assemble to form a functional tripartite Cas10, they may also have applications for type III CRISPR-based biotechnologies that could allow for selectable modularization of different Cas10 activities.

Cas10 cOA production requires all components of the type III surveillance complex

In addition to the differences in Cas10 domain architectures detailed above, Csm and Cmr complexes from different organisms have been reported to produce different ratios of cOA products. For example, Csm from *Streptococcus thermophilus* (SthCsm) mainly produces cA_6, while Csm from *Thermus thermophilus* (TtCsm) produces greater quantities of cA_4.[15,16] To determine if cOA-production profiles of Csm/Cmr complexes could be predicted based on phylogeny, we picked five representatives from three distinct clades of Cas10 evolution and tested the cOA-synthesis activity of these proteins. The five Cas10 homologs we purified all have GGDD motifs in their Palm2 domains and have been previously studied by other groups (i.e., TonCsm1, TthCsm1, SthCsm1, SsoCsm1, and PfuCmr2) (Figure 9.1B).[2,3,6,9,15,16,40,41]

To measure the efficiency of cOA synthesis, we incubated each Cas10 protein with P^{32}-labelled rATP for one hour. Products of this reaction were then resolved using thin layer chromatography (TLC). As positive controls, fully assembled TthCsm and SthCsm complexes were incubated with complementary RNAs and P^{32}-labelled rATP. As previously reported, fully assembled TthCsm and SthCsm produce cA_3, cA_4 and cA_6 when incubated with a

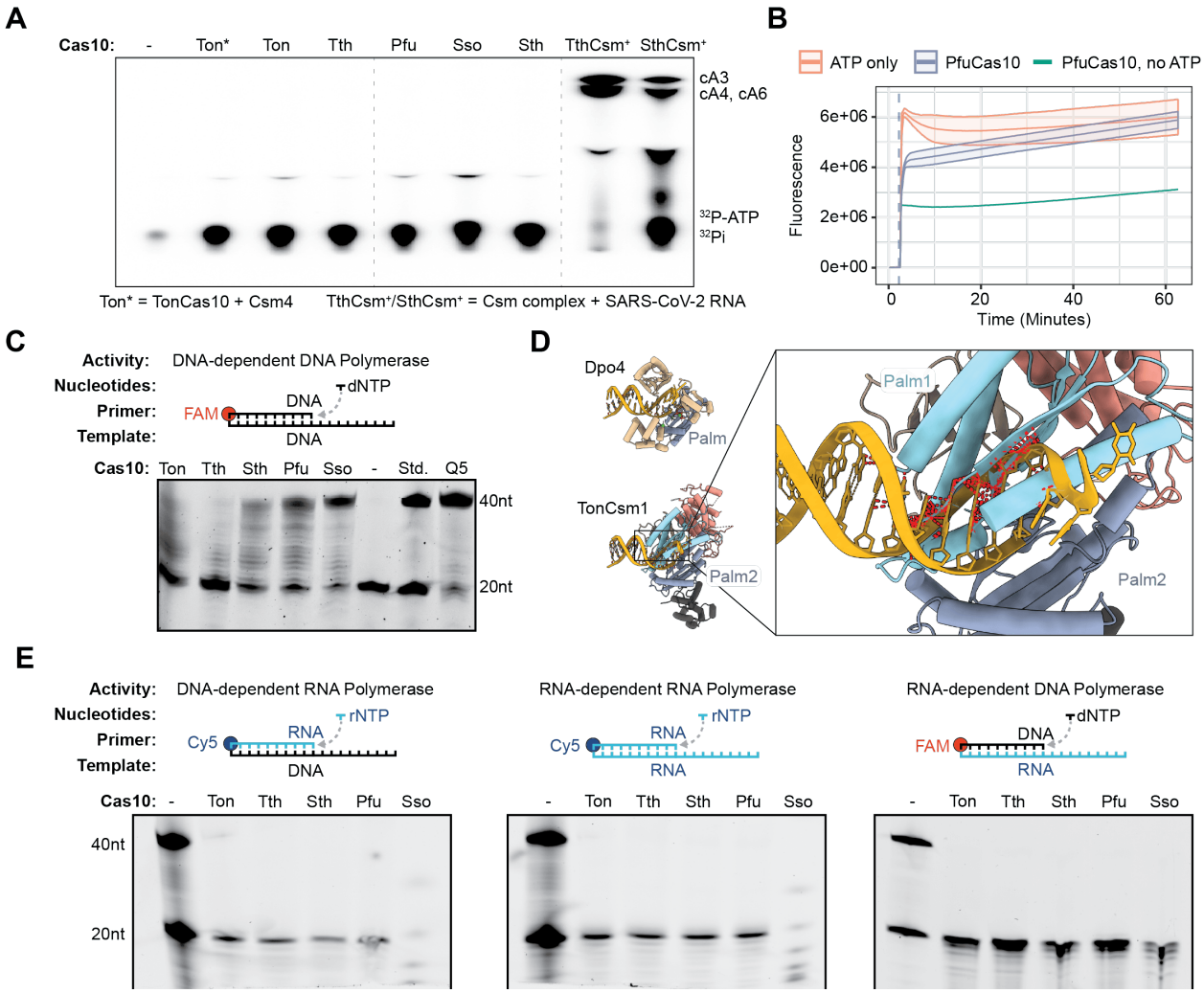


Figure 9.2. Cas10 alone is incapable of cOA-production but has DNA polymerase activity. (A) No P^{32}-labelled cOA products from Cas10 and Cas10-Csm4 are visible on a TLC plate. By contrast, cOA products corresponding to cA_3, cA_4, and cA_6 are visible when TthCsm or SthCsm complex are incubated with a crRNA-complementary target. (B) Representative fluorescence trace from Cas10 cyclase assay shows no differences in max fluorescence between reactions that contain both rATP and PfuCas10 and reactions that contain rATP alone. Middle line is shown as mean fluorescence values for three technical replicates, with a lower opacity window representing two standard deviations of each set of replicates. (Fluorescence traces for other Cas10 homologs tested in Figure S9.3.) (C) Fluorescently-labeled, higher molecular weight products (20-40nt) verify DNA-dependent DNA polymerase activity of five Cas10 homologs. 40nt and 20nt standards (Std.) are shown in lane seven for reference, and Q5 DNA polymerase positive control reaction products are shown in lane 8. (D) Left: to model Cas10 binding of a DNA primer-template substrate, a structure of TonCsm1 (PDB: 4UW2) was superimposed on a family Y DNA polymerase structure (Dpo4; PDB: 2AGQ) by aligning TonCsm1 Palm2 to Dpo4 Palm. Right: steric clashes between 704 DNA and Palm1 atoms are shown as red lines. (E) No

products corresponding to extended primers were detected when Cas10 proteins with RNA primers or RNA templates.

complementary RNA target (Figure 9.2A). In contrast, none of the Cas10 proteins—nor a heterodimer of TonCsm1 (Cas10) and TonCsm4—produced detectable levels of cOA (Figure 9.2A). These data indicate that additional Csm/Cmr components are needed to license Cas10 cOA-synthase activity (Figure 9.2A). To verify this finding, we independently tested polymerase activity using a Calcein reporter assay.[32,42] Similar to the TLC assay, no polymerase activity is detectable for any Cas10 protein in the presence of rNTPs, dNTPs, or rATP alone (Figures 9.2B and S9.1-S9.3). These results agree with recent experimental findings in the type III system from *Lactococcus lactis* (i.e., LlaCsm), showing that LlaCsm cOA-synthesis activity is dependent on target binding, since conformational changes in target-bound LlaCsm increase Cas10 protein dynamics.[43]

Diverse Cas10 proteins function as DNA-dependent DNA polymerases

Although we were unable to detect cOA formation for any of the five Cas10 proteins we tested, a recent report by Zhang et al. describes a template-dependent DNA polymerase activity for two of the Cas10 homologs we purified (i.e., TonCsm1 and PfuCmr2).[20] To verify the polymerase activity of these two Cas10s, and to assess the activity of the three additional Cas10 proteins we purified (i.e., SsoCsm1, TthCsm1, SthCsm1), we incubated each Cas10 with a 20 nucleotide FAM-labelled primer annealed to a 40 nucleotide template and deoxyribonucleotides (dNTPs). After one hour of incubation at 37ºC, products corresponding to extended primers were visible when reaction products were assessed on a denaturing UREA-PAGE gel (Figure 9.2C). These products are laddered in a manner that implies variable levels of primer extension,

possibly due to Cas10 dissociation from the primer-template substrate. While each Cas10 protein appeared to be capable of extending the DNA primer, the efficiencies of this reaction varied for each Cas10 homolog. Notably, TonCsm1, TthCsm1 and SthCsm1—all of the Cas10 proteins that clustered with III-A-associated Cas10s in our phylogenetic analysis—performed more poorly than PfuCmr2 (III-B) or SsoCsm1 (III-D) (Figure 9.2C). As a positive control, Q5 DNA polymerase was incubated with our FAM-labelled primer-template substrate and dNTPs for 10 minutes at 50ºC (Figure 9.2C). Taken together, these results suggest that DdDp polymerase activity may be a widely conserved feature of Cas10 proteins with Palm2 domains containing GGDD motifs.

To better understand how Cas10 functions as a DdDp, we superimposed structures of Cas10 from *Thermococcus onnurineus* (TonCsm1) onto the Palm domains of family Y DNA polymerases. Palm domains are common structural elements of reverse transcriptases, RNA-dependent RNA polymerases, family A, B and Y DNA polymerases, nucleotidyltransferases, nucleotidylcyclases, and Cas10 proteins.[19] A recent structural review of polymerases indicates that Palm2 from PfuCmr2 is most similar to Palm domains from family Y DNA polymerases.[18] When the Palm2 domain of TonCsm1—which contains the catalytic GGDD motif—is superimposed on the single Palm domain of Dpo4 from *Sulfolobus solfataricus* (i.e., a family Y DNA polymerase), the TonCsm1 Palm1 domain clashes with the double-stranded DNA substrate (Figure 9.2D). Similar clashing is also observed when Palm2 is superimposed on the Palm domains of other family Y DNA polymerases. These results suggest that the mechanism of Cas10 DNA-dependent DNA-polymerase activity is distinct from family Y DNA polymerases,

or else that Cas10 proteins undergo significant conformational rearrangements to accommodate DNA in the Palm2 active site.

Cas10-mediated DdDp activity utilizes deoxyribonucleotides (dNTPs), while cOA-synthesis by Cas10 requires ribonucleotides (i.e. rATP).[2,15] This led us to hypothesize that Cas10 proteins may function as more efficient polymerases of RNA than DNA. To test this hypothesis, we incubated each Cas10 with rNTPs and a Cy5-labelled RNA primer annealed to a DNA template. Surprisingly no products corresponding to extended primers were detectable in any replicate of this experiment (Figure 9.2E), indicating that none of the Cas10 homologs tested function as DNA-dependent RNA polymerases. Similarly, no products corresponding to extended primers were detectable with RNA-RNA or DNA-RNA primer-template substrates (Figure 9.2E), indicating that Cas10 has neither RNA-dependent RNA-polymerase nor reverse transcriptase activities, respectively. While it is unclear what biological role the DNA polymerase activity of Cas10 plays in CRISPR biology, the fact that this activity is conserved in five divergent Cas10 proteins suggests that it likely has some biological role. Additionally, it is worth noting that many polymerases require cofactors for efficient DNA synthesis, so some of the differences noted between the efficiencies of the five Cas10 homologs we tested could be due to the absence of additional cofactors that would be present in a cellular environment.[44-46]

Discussion

Collectively, this work shows that diversification of an ancestral Cas10 protein seeded the large subunits of type III and I-D surveillance complexes, with multiple independent losses of HD nuclease (in III-D, III-C, and III-B Cas10s) and Palm2-mediated cOA-synthesis activities (in III-F, III-C, and I-D) occurring over time. Indeed, our results indicate that roughly three-

quarters of all Cas10 proteins lack the residues that are expected to be required for HD nuclease activity. This work also puts the recently defined III-F systems into an evolutionary framework for the first time, revealing that III-F Cas10 proteins share a most recent common ancestor with Cas10s from III-C systems. The shared divergence of Palm2 domains in these two Cas10 subtypes suggests that cOA-signaling was lost in a common ancestor of III-F and III-C Cas10s. Notably, cOA-activated effectors (i.e., CARF-domain containing proteins) are absent in III-F loci, and we expect similar patterns for III-C loci.[5]

This work also highlights the modularity of CRISPR-Cas components by showing that reshuffling and novel fusions of protein domains has contributed to the diversification of bacterial and archaeal immune systems. We identified two novel Cas10 fusions that represent the incorporation of a cOA-gated effector into the crRNA-guided surveillance complex (NucC-Cas10) and the integration of a toxin-antitoxin system into type III-C CRISPR systems (AbiEii-Cas10). Additionally, we identified a system in which Cas10 is split over multiple genes (i.e., III-A in *T. marina*). These examples support the 'guns for hire' theory, and demonstrate how CRISPR systems become diversified over time due to reshuffling of immune and non-immune components.[47] The linkage of CRISPR-Cas with a toxin-antitoxin system—in the examples of AbiEii-Cas10—also supports the hypothesis that immune components are often spread by mobile genetic elements that have a vested self-interest in keeping these systems intact. Indeed, many CRISPR-Cas systems are encoded on mobile genetic elements (e.g., plasmids and transposons) and a recent report details a Cas-mediated toxin-antitoxin system that penalizes the loss of certain I-E CRISPR-Cas components.[48] AbiEii may have once similarly been a separate

component of a type III-A locus, that prevented degeneration or loss of the system, before it was genetically fused to Cas10.

Our biochemical work reveals that GGDD-motif containing Cas10 proteins alone are insufficient for cyclase activity, but instead these proteins function as DNA-dependent DNA polymerases (DdDp). The lack of cOA-synthesis activity seen from Cas10 proteins—without other Csm/Cmr components and a crRNA-complementary target—is unsurprising when one considers that cyclase activity is likely to be tightly regulated in a cellular environment. Promiscuous cOA-production would lead to activation of cOA-gated effectors, inducing cellular dormancy or cell-death in the absence of bacteriophage infection. More surprising, however, is our data verifying the DdDp activity of five diverse Cas10 homologs, since it suggests that this activity is likely a widely conserved feature of Cas10 proteins with intact Palm2 GGDD motifs. Small DdDp polymerases, which are unrelated to Cas10 and are called CRISPR-Associated Primase-Polymerases (CAPPs), were recently reported to exist in *cas* loci.[49] CAPPs are hypothesized to aid in the production of prespacers during CRISPR adaptation, suggesting a possible analogous role for Cas10 DdDp activity. Cas10 polymerase activity is confined to extension of DNA primers using DNA templates and requires deoxyribonucleotides, rather than ribonucleotides (i.e., the primary ligand during cOA-synthesis). Crystal structures of TonCsm1 and Csm4 bound to multiple ligands reveal that TonCsm1 is able to bind deoxyribonucleotides using the same Palm domain residues used to coordinate ribonucleotides.[2] However, our structural modeling raises questions about how a DNA primer and DNA template are coordinated by Cas10 during polymerization, since the Palm1 domain of Cas10 prevents positioning of these substrates at the active site in a manner that is analogous to family Y DNA

polymerases. Structural determination of Cas10 bound to a DNA-primer DNA-template substrate could resolve the mechanism of Cas10 polymerase activity and may help explain the differences in DdDp efficiency we observed between five Cas10 homologs.

Acknowledgements

B.W. is the founder of SurGene, LLC, and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. Research in the Wiedenheft lab is supported by the National Institutes of Health (R35GM134867) and the Montana State University Agricultural Experimental Station (USDA NIFA).

References

1 Huo, Y. *et al.* Cryo-EM structure of Type III-A CRISPR effector complex. *Cell Res* **28**, 1195-1197, doi:10.1038/s41422-018-0115-6 (2018).

2 Jia, N., Jones, R., Sukenick, G. & Patel, D. J. Second Messenger cA4 Formation within the Composite Csm1 Palm Pocket of Type III-A CRISPR-Cas Csm Complex and Its Release Path. *Mol Cell* **75**, 933-943 e936, doi:10.1016/j.molcel.2019.06.013 (2019).

3 Liu, T. Y., Liu, J. J., Aditham, A. J., Nogales, E. & Doudna, J. A. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. *Nat Commun* **10**, 3001, doi:10.1038/s41467-019-10780-2 (2019).

4 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).

5 Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, doi:10.1038/s41579-019-0299-x (2019).

6 Sofos, N. *et al.* Structures of the Cmr-beta Complex Reveal the Regulation of the Immunity Mechanism of Type III-B CRISPR-Cas. *Mol Cell* **79**, 741-757 e747, doi:10.1016/j.molcel.2020.07.008 (2020).

7 Ozcan, A. *et al.* Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. *Nature* **597**, 720-725, doi:10.1038/s41586-021-03886-5 (2021).

8 Manav, M. C. *et al.* Structural basis for inhibition of an archaeal CRISPR-Cas type I-D large subunit by an anti-CRISPR protein. *Nat Commun* **11**, 5993, doi:10.1038/s41467-020-19847-x (2020).

9 Jung, T. Y. *et al.* Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* **23**, 782-790, doi:10.1016/j.str.2015.01.021 (2015).

10 Elmore, J. R. *et al.* Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev* **30**, 447-459, doi:10.1101/gad.272153.115 (2016).

11 Estrella, M. A., Kuo, F. T. & Bailey, S. RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes Dev* **30**, 460-470, doi:10.1101/gad.273722.115 (2016).

12 Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, C. & Siksnys, V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Mol Cell* **62**, 295-306, doi:10.1016/j.molcel.2016.03.024 (2016).

13 Mo, C. Y. *et al.* Type III-A CRISPR immunity promotes mutagenesis of staphylococci. *Nature* **592**, 611-615, doi:10.1038/s41586-021-03440-3 (2021).

14 Hatoum-Aslan, A., Maniv, I., Samai, P. & Marraffini, L. A. Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. *J Bacteriol* **196**, 310-317, doi:10.1128/JB.01130-13 (2014).

15 Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609, doi:10.1126/science.aao0100 (2017).

16 Niewoehner, O. *et al.* Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548, doi:10.1038/nature23467 (2017).

17 Athukoralage, J. S. & White, M. F. Cyclic oligoadenylate signalling and regulation by ring nucleases during type III CRISPR defence. *RNA*, doi:10.1261/rna.078739.121 (2021).

18 Monttinen, H. A., Ravantti, J. J. & Poranen, M. M. Common Structural Core of Three-Dozen Residues Reveals Intersuperfamily Relationships. *Mol Biol Evol* **33**, 1697-1710, doi:10.1093/molbev/msw047 (2016).

19 Yang, W. An overview of Y-Family DNA polymerases and a case study of human DNA polymerase eta. *Biochemistry* **53**, 2793-2803, doi:10.1021/bi500019s (2014).

20 Zhang, S. *et al.* Mycobacterium tuberculosis CRISPR/Cas system Csm1 holds clues to the evolutionary relationship between DNA polymerase and cyclase activity. *Int J Biol Macromol* **170**, 140-149, doi:10.1016/j.ijbiomac.2020.12.014 (2021).

21 McKay, L. J., Hatzenpichler, R., Inskeep, W. P. & Fields, M. W. Occurrence and expression of novel methyl-coenzyme M reductase gene (mcrA) variants in hot spring sediments. *Sci Rep* **7**, 7252, doi:10.1038/s41598-017-07354-x (2017).

22 Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* **46**, W246-W251, doi:10.1093/nar/gky425 (2018).

23 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

24 Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-682, doi:10.1093/bioinformatics/btq003 (2010).

25 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

26 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518, doi:10.1093/nar/gki198 (2005).

27 Yu, G., Lam, T. T., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**, 3041-3043, doi:10.1093/molbev/msy194 (2018).

28 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

29 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419, doi:10.1093/nar/gkaa913 (2021).

30 Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14-25, doi:10.1002/pro.3235 (2018).

31 Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322, doi:10.1186/s12859-015-0749-z (2015).

32 Santiago-Frangos, A. *et al.* Intrinsic signal amplification by type III CRISPR-Cas systems provides a sequence-specific SARS-CoV-2 diagnostic. *Cell Rep Med* **2**, 100319, doi:10.1016/j.xcrm.2021.100319 (2021).

33 Lin, J. *et al.* DNA targeting by subtype I-D CRISPR-Cas shows type I and type III features. *Nucleic Acids Res* **48**, 10470-10478, doi:10.1093/nar/gkaa749 (2020).

34 Rouillon, C., Athukoralage, J. S., Graham, S., Gruschow, S. & White, M. F. Control of cyclic oligoadenylate synthesis in a type III CRISPR system. *Elife* **7**, doi:10.7554/eLife.36734 (2018).

35 Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* **21**, 771-777, doi:10.1038/nsmb.2875 (2014).

36 Beloglazova, N. *et al.* Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J* **30**, 4616-4627, doi:10.1038/emboj.2011.377 (2011).

37 Lau, R. K. *et al.* Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity. *Mol Cell* **77**, 723-733 e726, doi:10.1016/j.molcel.2019.12.010 (2020).

38 Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. *Nucleic Acids Res* **42**, 4590-4605, doi:10.1093/nar/gkt1419 (2014).

39 Athukoralage, J. S. *et al.* An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* **577**, 572-575, doi:10.1038/s41586-019-1909-5 (2020).

40 Cocozaki, A. I. *et al.* Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure* **20**, 545-553, doi:10.1016/j.str.2012.01.018 (2012).

41 Osawa, T., Inanaga, H. & Numata, T. Crystal structure of the Cmr2-Cmr3 subcomplex in the CRISPR-Cas RNA silencing effector complex. *J Mol Biol* **425**, 3811-3823, doi:10.1016/j.jmb.2013.03.042 (2013).

42 Tomita, N., Mori, Y., Kanda, H. & Notomi, T. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nat Protoc* **3**, 877-882, doi:10.1038/nprot.2008.57 (2008).

43 Sridhara, S. *et al.* Structure and Function of an in vivo Assembled Type III-A CRISPR-Cas Complex Reveal Critical Roles of Dynamics in Activity Control. *bioRxiv*, 2021.2001.2027.428455, doi:10.1101/2021.01.27.428455 (2021).

44 Casteel, D. E. *et al.* A DNA polymerase-{alpha}{middle dot}primase cofactor with homology to replication protein A-32 regulates DNA replication in mammalian cells. *J Biol Chem* **284**, 5807-5818, doi:10.1074/jbc.M807593200 (2009).

45 Kumble, K. D., Iversen, P. L. & Vishwanatha, J. K. The role of primer recognition proteins in DNA replication: inhibition of cellular proliferation by antisense oligodeoxyribonucleotides. *J Cell Sci* **101 (Pt 1)**, 35-41 (1992).

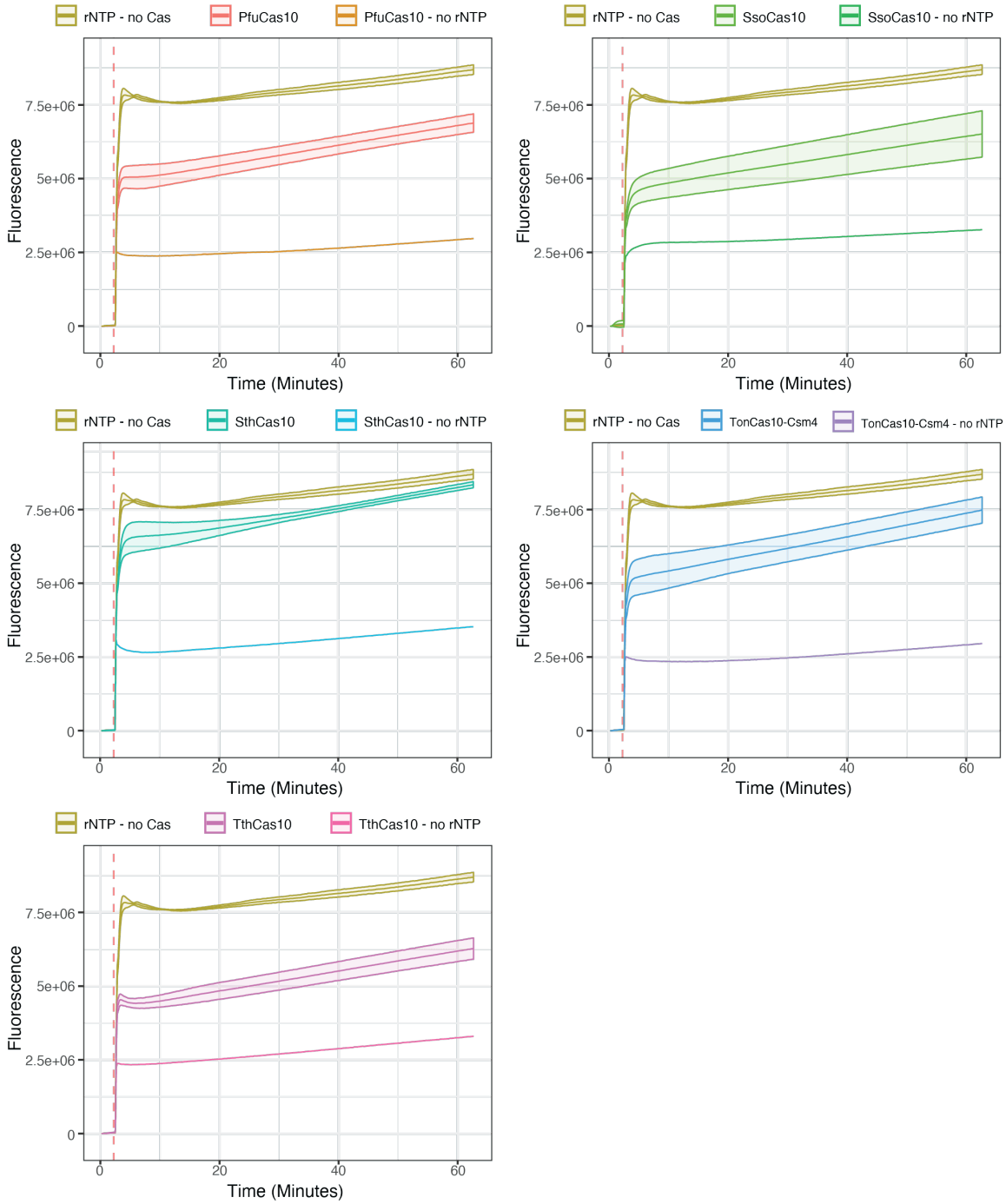
46 Pritchard, C. G., Weaver, D. T., Baril, E. F. & DePamphilis, M. L. DNA polymerase alpha cofactors C1C2 function as primer recognition proteins. *J Biol Chem* **258**, 9810-9819 (1983).

47 Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet* **21**, 119-131, doi:10.1038/s41576-019-0172-9 (2020).

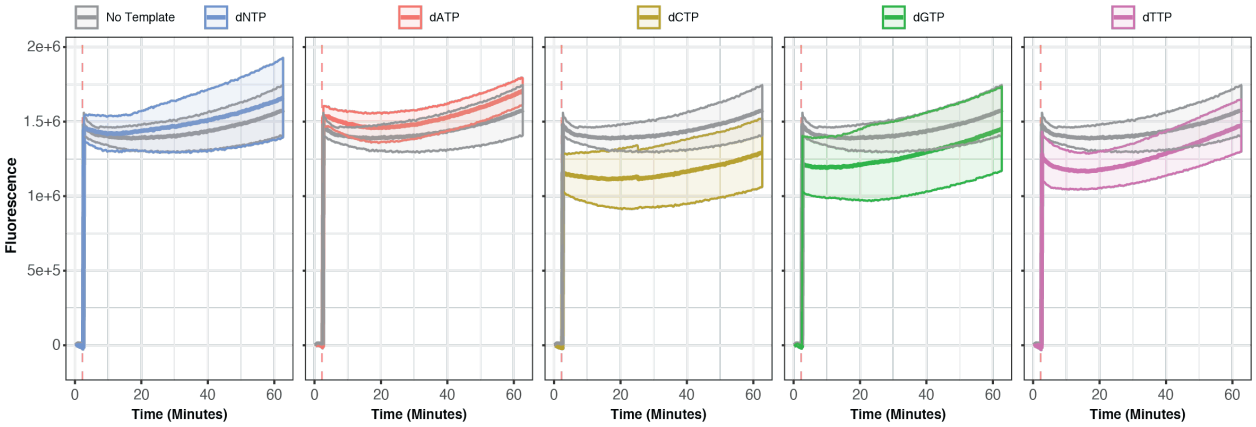
48 Li, M. *et al.* Toxin-antitoxin RNA pairs safeguard CRISPR-Cas systems. *Science* **372**, doi:10.1126/science.abe5601 (2021).

49 Zabrady, K., Zabrady, M., Kolesar, P., Li, A. W. H. & Doherty, A. J. CRISPR-Associated Primase-Polymerases are implicated in prokaryotic CRISPR-Cas adaptation. *Nat Commun* **12**, 3690, doi:10.1038/s41467-021-23535-9 (2021).

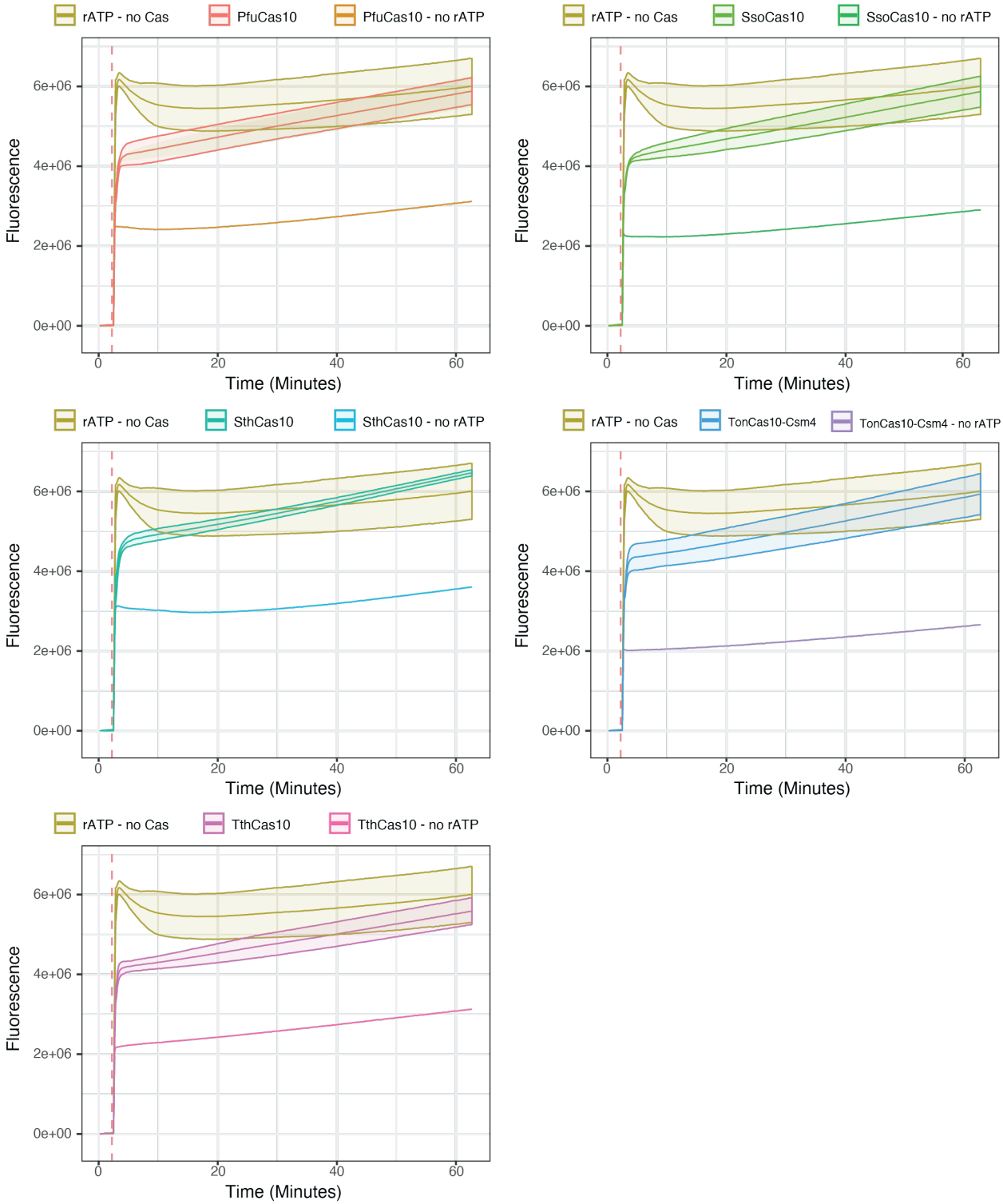
Supplemental Figures



Supplemental Figure S9.1. Fluorescence traces of Calcein assays to test cyclase activity with rNTPs. No cyclase activity is detectable for any of the Cas10 proteins tested. Data are shown as the mean of three technical replicates, with two standards of deviation shown as shading of the same color. Negative control reactions do not contain Cas10 and are labelled as 'no Cas'.



Supplemental Figure S9.2. Fluorescence traces of Calcein assays to test PfuCmr2 cyclase activity with dNTPs. No cyclase activity is detectable for PfuCmr2. Data are shown as the mean of three technical replicates, with one standard of deviation shown as shading of the same color. The 'dNTP' condition was supplied with an equimolar mix of dATP, dCTP, dGTP, and dTTP.



Supplemental Figure S9.3. Fluorescence traces of Calcein assays to test cyclase activity with rATP only. No cyclase activity is detectable for any of the Cas10 proteins tested. Data are shown as the mean of three technical replicates, with two standards of deviation shown as shading of the same color. Negative control reactions do not contain Cas10 and are labelled as 'no Cas'.

Supplemental Tables

		HD2 Motif		HD1 Motif		GGDD (Palm2) Motif	
Subtype	**Total**	N	Percent	N	Percent	N	Percent
III-A	2715	842	31.01	946	34.84	1712	63.06
III-B	2736	0	0	0	0	1868	68.27
III-C	666	0	0	52	7.81	0	0
III-D	1506	0	0	0	0	1073	71.25
III-F	144	0	0	50	34.72	0	0
I-D	321	0	0	192	59.81	0	0

Supplemental Table S9.1. Quantification of HD and Palm2 catalytic site motifs for different Cas10 subtypes.

CHAPTER TEN

PHYLOGENETIC ANALYSIS OF CAS7 REVEALS GENETIC FUSIONS TO TOXIN-ANTITOXIN AND INNATE IMMUNITY COMPONENTS

Contribution of Authors and Co-Authors

Manuscript in Chapter 10

Author: Tanner Wiegand

Contributions: designed the research, performed the research, prepared manuscript figures, wrote the manuscript, wrote the manuscript

Co-Author: Mackenzie Lynes

Contributions: acquired samples and performed metagenomic sequencing

Co-Author: Roland Hatzenpichler

Contributions: acquired samples and performed metagenomic sequencing

Co-Author: Blake Wiedenheft

Contributions: designed the research, wrote the manuscript

Manuscript Information

Tanner Wiegand, Mackenzie Lynes, Roland Hatzenpichler, Blake Wiedenheft

The CRISPR Journal

Status of Manuscript: [Put an x in one of the options below, delete this]
__X_ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
____ Published in a peer-reviewed journal

276

PHYLOGENETIC ANALYSIS OF CAS7 REVEALS GENETIC FUSIONS TO TOXIN-ANTITOXIN AND INNATE IMMUNITY COMPONENTS

Tanner Wiegand,[1] Mackenzie Lynes,[2] Roland Hatzenpichler,[2] and Blake Wiedenheft[1]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA. [2]Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717, USA.

Abstract

Cas7 proteins are conserved components of the type I, III, and IV CRISPR (clustered regularly interspaced short palindromic repeats) RNA-guided surveillance complex. Here, we identify 22,983 unique Cas7 proteins in genomic and metagenomic databases and show that Cas7 proteins from type III systems are likely to have a more ancient origin than Cas7 proteins from type I and type IV systems. We additionally identify seven examples of Cas7 proteins that are genetically fused to other immune (e.g., CBASS) or toxin-antitoxin components. We discuss how these fusions highlight the mobile and plastic nature of CRISPR-Cas components and predict how these newly identified Cas7 fusions are likely to function in CRISPR-mediated immunity.

Introduction

In class I CRISPR systems, multiple protein subunits assemble with a CRISPR RNA (crRNA) to form a surveillance complex.[1,2] Type I, III, and IV CRISPR systems—encompassing fifteen CRISPR subtypes—incorporate at least one Cas7 homolog into the complex.[3-11] Cas7 proteins have been referred to by many different names (Table S10.1). In type I systems, Cas7 proteins are generally appended with a letter corresponding to their associated subtype (e.g., Cas7e proteins are found in I-E systems).[1,2] Type IV Cas7s are called Csf2, since they were originally described as CRISPR-Cas systems of *A. ferroxidans*.[12,13] Things become more complicated with type III systems—since multiple Cas7 homologs are incorporated into type III Cascade complexes—prompting the need for various names including Csm3 and Csm5 for type III-A and III-D systems and Cmr1, Cmr4, and Cmr6 for type III-B and III-C systems.[1,2,13] Cas7 proteins are often referred to as Cascade 'backbone' subunits, since three to five Cas7s typically bind to the central region of a crRNA.[3-11] Cmr1 and Cmr6 break this paradigm, binding as single copies to the 3' end of a crRNA guide.

Here, we use profile Hidden Markov Models (HMMs) to identify 22,983 unique Cas7 proteins in genomic and metagenomic databases. We show that Cas7 proteins from type I and type IV systems likely emerged from a more recent common ancestor than Cas7 proteins from type III systems. This work supports hypotheses that type I and IV systems evolved from the diversification of type III systems.[1,2,14] Our results show that Cas7 proteins have been genetically fused to components from other immune systems and toxin-antitoxin systems. Some of the fusions that we identify (e.g., Cas7c-TniQ, Cas3-Cas7e, Cas10-Csm3) are likely to be integrated in multiple copies in a single crRNA-guided surveillance complex, raising mechanistic questions

about how these Cas7 fusions function in immunity. Other Cas7 fusions we identify (e.g., Cmr6-TIR-SAVED and Cmr1-MntA) highlight repurposing of alternate immune and selfish genetic element components by CRISPR-Cas systems, and indicate that there are variant class I CRISPR systems that still await further investigation.

Materials and Methods

Cas7 identification

Prodigal was used to translate all open reading frames (ORFs) in 58,864 genomes downloaded from the NCBI database, 9,687 genomes from Joint Genomics Institute (JGI) and 26,354,826 metagenomic scaffolds from samples isolated in Yellowstone National Park.[15] Profile hidden markov models (HMMs) of Cas7 were pulled from CasFinder and used to query the entire database (HMMs described in Table S10.1).[16] HMM queries were conducted in hmmer with an E-value cutoff of 0.01, before search results were filtered in R to exclude matches that did not meet a sequence E-value threshold of 1e-05.[17] This search resulted in the identification of 33,964 Cas7-like sequences. Redundant sequences were filtered using CD-HIT, resulting in a set of 22,983 Cas7-like proteins.[18]

Phylogenetic and domain architecture analyses

22,983 unique Cas7-like sequences were aligned in MAFFT using the “auto” setting, and columns comprised of more than 30% gaps were removed from resulting alignment using Trimal.[19,20] The alignment was then used to construct a phylogenetic tree in FastTree with -wag and -gamma model settings.[21] Phylogenetic trees were visualized using the ggTree package from BioConductor in R, after being rerooted to the midpoint of the tree using the Phangorn

package.[22,23] Each HMM used for initial Cas7 identification was composed of subtype specific Cas7 protein sequences (e.g., III-A Csm3s only), which enabled subtypes to be assigned to each Cas7-like sequence. To determine if ancillary domains were present in Cas7-like sequences, we used the PFAM database and profile HMMs of Cas proteins extracted from the CasFinder program to annotate each Cas7-like sequence.[16,24] Results were manually analyzed using HHpred and structural predictions were generated using Phyre2 and the Alphafold "predict" utility in ChimeraX.[25-28]

Results

Cas7 phylogeny supports the proposed recent origin of type I systems

To investigate the phylogenetic diversity of Cas7 proteins, we used profile hidden markov models (HMMs) built from alignments of Cas7 to query the NCBI and JGI genomic databases. Additionally, we queried a set of 26,354,826 metagenomic scaffolds sequenced from environmental samples obtained in Yellowstone National Park. This search resulted in the identification of 22,983 unique Cas7 protein sequences, which were aligned and used to construct a phylogenetic tree.

Cas7 proteins from type I and type IV systems (i.e., I-A, -B, -C, and -E) form monophyletic clades near the head of the tree (Figure 10.1A). Notably, Cas7 proteins from type I-B and type I-C systems share the most recent common ancestor in our tree, indicating these proteins emerged more recently than Cas7s from other class I CRISPR systems. This clustering, and the presence of a shared node upstream, also suggests that Cas7 proteins from type I systems share a recent common ancestor and may have emerged more recently than Cas7 proteins from type III systems. These data agree with our analyses of Cas10 proteins, which suggests that

components of type III CRISPR systems evolved from more distant ancestors than components of type I systems (see Chapter 9). This evolutionary path (i.e., the emergence of type I systems via diversification of type III CRISPR-Cas) was first proposed by Eugene Koonin's group and is supported by phylogenetic analyses of Cas1 adaptation proteins.[1,2,14]

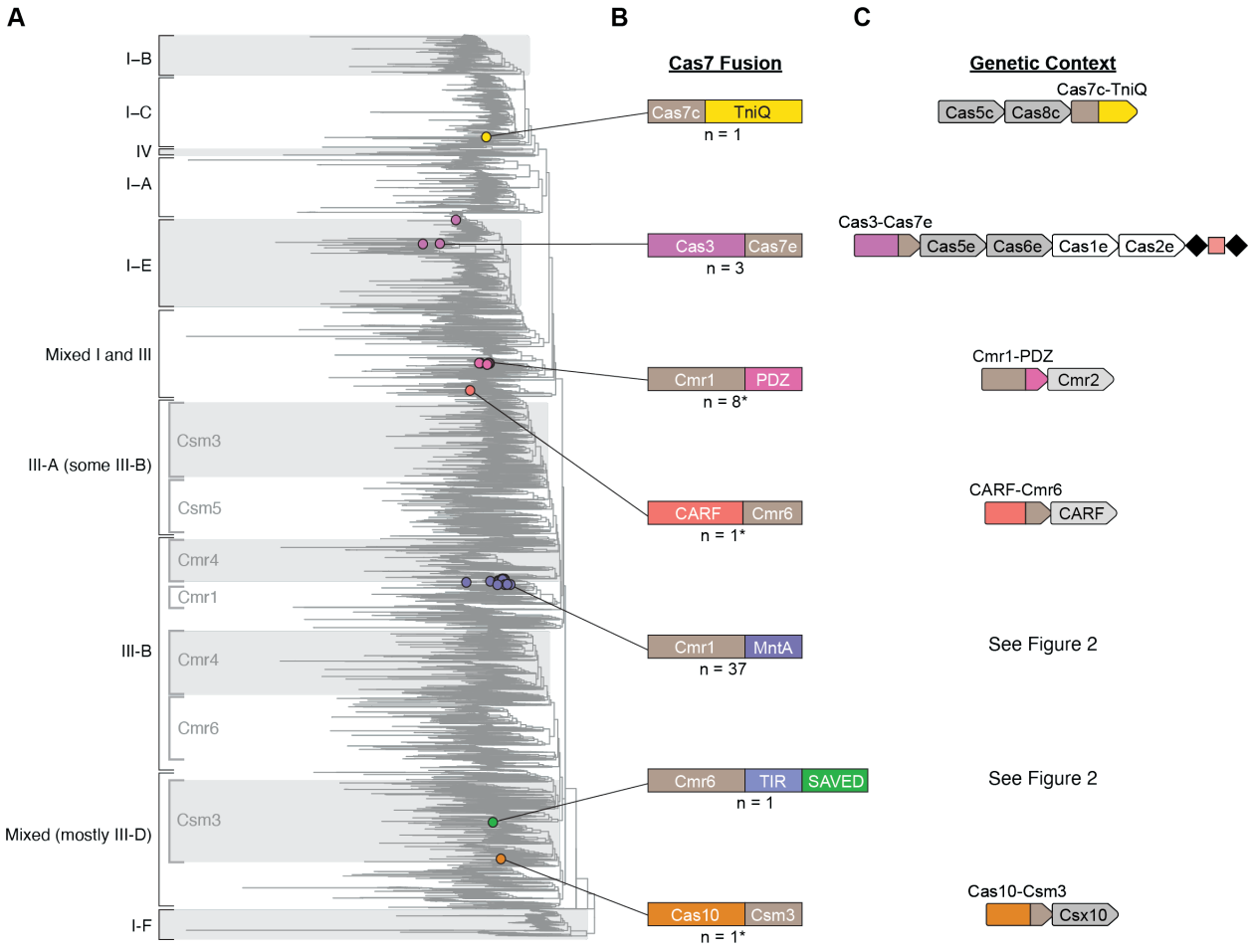


Figure 10.1. Phylogenetic analysis of Cas7 sequences reveals novel genetic fusions. (A) Maximum-likelihood phylogenetic tree built from multiple alignment of 22,983 unique Cas7 protein sequences. Newly identified Cas7 fusions are shown as points on the tree. (B) Schematic depictions of protein domain organization in novel Cas7 fusion proteins. Number of examples of each fusion identified are shown under each schematic. Asterisks indicate that all examples were identified on short metagenomic scaffolds, preventing comprehensive analysis of a fusion's encoding locus. (C) Genetic locus schematized for one representative of each Cas7 fusion.

Interestingly, type I-F-associated Cas7 sequences do not fall into a clade that is situated near I-E Cas7s, as we suspected they would due to the close relationship between I-F and I-E Cas1 proteins.[1,13] Rather, I-F Cas7 proteins are separated into their own clade that emerges at the trifurcated root of the tree (Figure 10.1A), indicating that our phylogenetic approach was unable to place I-F Cas7s into an evolutionary context with Cas7s from other CRISPR subtypes. While these data may suggest that Cas7f proteins are strongly diverged at the primary sequence level, structural studies of Cas7 proteins in I-E and I-F Cascade complexes indicate that tertiary structure is conserved in Cas7f proteins.[3,6]

Cas7 proteins are genetically fused to other Cas and CRISPR-associated transposition proteins

To assess the domain architectures of Cas7 proteins, we annotated each sequence using HMMs from the CasFinder program and the PFAM database.[16,24] This analysis revealed three examples of Cas7 proteins that have been genetically fused to other Cas proteins. Additionally, we identified Cas7 fusions that include components of alternate immune systems and toxin-antitoxin systems. One prevalent example that did not fall into these categories were eight type III Cas7 proteins (i.e., Cmr1) that are fused at their C-terminus to a domain called PDZ that is often involved in cytosolic localization and cytoskeletal binding (Figure 10.1B).[29] In all, we noted seven Cas7 fusions—totaling 52 sequences—that warranted further investigation (Figure 10.1B).

One of the fusions we identified was a Cas7c protein with a TniQ domain (Figure 10.1B). Tn7-like transposons use TniQ—in concert with other transposition machinery (e.g. TnsA, TnsB, and TnsC)—to integrate themselves into bacterial genomes.[30] Numerous reports have detailed Tn7-like transposons that repurpose CRISPR systems to direct their TniQ-mediated mobilization

(i.e., CRISPR-associated transposons; CASTs), including a recent pre-print describing the identification of CASTs that utilize a I-C Cascade.[31-36] While the genetic fusion of Cas7c-TniQ has not been previously identified, TniQ is fused to Cas5 or Cas8 proteins in some I-F CAST systems, suggesting that TniQ has been incorporated as a genetically fused subunit of Cascade on multiple occasions.[35] The fusion of TniQ to Cas7c is particularly interesting, since Cas7c is expected be present in seven copies along the crRNA backbone of I-C Cascade.[8] It is unclear how this stoichiometry of TniQ-containing proteins impacts transposition in this variant I-C CAST system, since other CAST systems use only one or two copies of TniQ to direct the transposition complex to a target site.[31,37,38]

In addition to CAST proteins, we identified Cas7 sequences with HD domains resulting from genetic fusions with Cas3 or Cas10 (Figure 10.1B). Cas3 is a nuclease-helicase that is recruited to degrade DNA targets bound by type I Cascades, and there were three examples of Cas3 fused to the N-terminus of Cas7e in our dataset.[39-43] Based on structures of I-E Cascade from *E. coli*, we expect that four copies of Cas3-Cas7e will be incorporated into the surveillance complex, raising questions about which Cas3-Cas7e subunits are actually responsible for target degradation in these systems.[6] It is worth noting that Cas3 is fused to Cas2 in I-F systems (i.e. Cas2-3), and Cas3 proteins are thought to have evolved from the HD domains of Cas10 proteins (a large subunit of type III Cascade) (see Chapter 9).[13,14] Interestingly, we identified one example of a type III-D Cas7 protein (Csm3) that is fused at its N-terminus to a Cas10 HD and partial Palm1 domain (Figure 10.1B). Three to five copies of Cas10-Csm3 are expected to be present in the variant III-D Cascade from this system. This Cas10-Csm3 fusion is conceptually similar to

the Cas3-Cas7e fusions, since it would result in the incorporation of an HD domain into multiple backbone subunits of Cascade.

Finally, we noted a type III-B Cas7 protein (i.e., Cmr6) that is genetically fused at its N-terminus to a cyclic-oligoadenylate binding (i.e., CARF) domain (Figure 10.1B). CARF domains regulate effector proteins in type III CRISPR systems (e.g., Csx1, Can1, Card1 etc.), activating their nuclease activities only when cyclic-oligoadenylate (cOA) molecules are bound to two dimerized CARF domains.[44-47] It is interesting to note the genetic context of the CARF-Cmr6 fusion in our dataset, since the gene immediately downstream encodes another protein that is predicted to be comprised of only a CARF domain (Figure 10.1B). It is conceivable Cmr6-CARF forms a heterodimer with the neighboring CARF protein, although the function of this interaction would not be immediately apparent. Unfortunately, the single example of Cmr6-CARF we identified was encoded in a short metagenomic scaffold, so we were unable to further assess the genetic context of this protein for additional clues to its possible role in CRISPR defense.

Fusion of Cas7 to components of a HEPN-based toxin-antitoxin system

Our search for novel Cas7 fusions revealed 37 examples of Cmr1 with a C-terminal nucleotidyltransferase domain (Figure 10.1B). Closer inspection of these proteins revealed them to be fusions of Cmr1 with MntA, an antitoxin that prevents activation of a HEPN RNAse (i.e., HepT) via ampylation of the toxin's active site.[48,49] Cmr1-MntA is encoded upstream of a HepT homolog and type III-B Cascade components (i.e., Cmr2 through Cmr6 and a CRISPR) (Figure 10.2A).

To further assess the possible role of Cmr1-MntA fusions, we focused our analysis on the variant III-B locus from *Dissulfurispira thermophila* (AP022873.1; Genbank) (Figure 10.2A). To

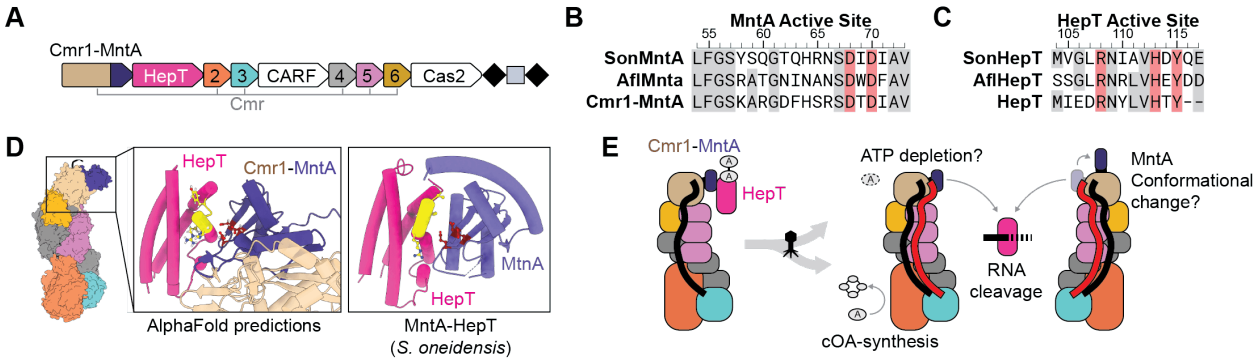


Figure 10.2. Genetic fusion of Cas7 to an antitoxin suggests repurposing of TA system components for CRISPR immunity. (A) Map of Cmr1-MntA genetic locus from *Dissulfurispira thermophila* (AP022873.1; Genbank) includes Cmr components (Cmr proteins and a CRISPR) and a HEPN toxin (Hept). (B) Two aspartic acid residues in MntA homologs from *Shewanella oneidensis* (SonMntA) and *Aphanizomenon flos-aquae* (AflMntA) that facilitate ampylation of HepT are conserved in the active site of Cmr1-MntA. (C) Three residues that catalyze RNA cleavage in *S. oneidensis* (SonHepT) and *A. flos-aquae* HepT homologs are conserved in Cmr1-MntA. (D) Left: Cartoon depiction of Alphafold-predicted Cmr1-MntA structure superimposed on the Cmr1 subunit of surveillance complex from *Sulfolobus islandicus* (PDB: 6S6B) reveals that one copy of Cmr1-MntA is likely to be present in variant III-B CascadeCascade. Cmr subunits are colored as in panel A. Alphafold-predicted structures of Cmr1-MntA and associated HepT homolog (left inset) superimposed on structures of *S. oneidensis* HepT-MntA complex (right inset; PDB: 6M6U) suggest that residues necessary for HepT ampylation are accessible on Cmr1-MntA. (E) Schematic of predicted Cmr1-MntA role in CRISPR-mediated immunity. Target RNA binding by variant III-B CascadeCascade licenses cOA-formation by Cas10, depleting the rATP pool required to ampylate HepT (center). Alternatively, target binding by variant III-B CascadeCascade induces conformational changes in MntA that prevent ampylation of HepT (right). Either mechanism could result in HepT RNAse activity, prompting cell death or dormancy

determine if the MntA domain of this fusion—and the genetically neighboring HepT protein from this system—are likely to function like other reported MntA-HepT systems, we compared the sequence of each protein to MntA and HepT sequences from *Shewanella oneidensis* and *Aphanizomenon flos-aquae*.[48,49] Multiple sequence alignments revealed that Cmr1-MntA and its associated HepT protein each contain the conserved active-site residues that are expected to facilitate toxin-ampylation and RNA cleavage, respectively (Figure 10.2B and C).

To better understand how Cmr1-MntA is likely to be incorporated into III-B Cascade, we superimposed an Alphafold-predicted structure of this fusion on the Cmr1 subunit of III-B Cascade from *Sulfolobus islandicus* (SisCmr).[10] Cmr1 is present as a single copy in type III-B Cascade, binding to the 3'-end of the crRNA at the head of the complex (Figure 10.2D). This suggests that a single copy of Cmr1-MntA is therefore likely to be present in the variant III-B Cascade. Notably, when the predicted structure of Cmr1-MntA is superimposed on SisCmr1, the active site residues in the MntA domain are externally presented in a manner that suggests they would be accessible to ampylate HepT (Figure 10.2D). Cmr1-MntA lacks a C-terminal alpha-helix that is present in structures of MntA from *S. oneidensis* (Figure 10.2D), however this alpha-helix is truncated in many naturally occurring MntA proteins, so it is unlikely to be required for HepT ampylation in this system.[48] Considered together with the conservation of Cmr1-MntA and HepT active site motifs, these observations suggest that this variant III-B CRISPR system may have Cmr1-MntA-mediated ampylation and HepT-mediated nuclease activities.

Based on the reported mechanisms of other MntA-HEPN TA systems, we predict that the toxicity of HepT in these variant III-B systems may have been coopted for immune purposes. HepT RNAses are normally inactivated due to ampylation by MntA, but they become activated when cellular ATP concentrations become depleted, licensing cleavage of either mRNAs or tRNAs.[48,49] Cmr1-MntA may similarly keep the activity of HepT at bay until bacteriophage RNA is detected by the variant III-B Cascade. In many type III systems, the detection of crRNA-complementary targets triggers cOA synthesis by the Cas10 subunit of Cascade.[44,46,47] ATP is the primary ligand Cas10 uses to produce cOA molecules, but ATP is also required for MntA-mediated ampylation of HepT.[48] Thus, it is conceivable that ATP is depleted by Cas10 when a

variant III-B Cascade binds an RNA target, preventing further ampylation of HepT by Cmr1-MntA (Figure 10.2E). Alternatively, target binding may induce conformational changes in variant III-B Cascade that prevent interactions between Cmr1-MntA and HepT (Figure 10.2E). Regardless of what mechanism these systems utilize, the result is likely to be the same: HepT activity would no longer be repressed and cleavage of mRNAs or tRNAs would induce an abortive infection or cellular dormancy. This predicted model seems particularly compelling in light of the fact that many type III CRISPR effectors halt bacteriophage infection by indiscriminately cleaving RNA (e.g. Csx1 and Csm6).[45] Many other innate prokaryotic immune systems also stop phage replication, thereby protecting neighboring cells, via induced dormancy or programmed cell death mechanisms (e.g. cBASS).[50-53]

Cas7 proteins are fused to components of innate immune systems

A final interesting Cas7 fusion that we identified is composed of an N-terminal Cmr6 domain with a TIR and C-terminal SAVED domain (i.e., Cmr6-TIR-SAVED) (Figure 10.1B). TIR (toll-interleukin-1 receptor) domains are incorporated in immune signaling proteins from all domains of life and have $NADP^+$ cleavage activity.[54] SAVED domains are similarly found in innate immune proteins of bacteria—where they act as receptors of cyclic-oligonucleotide second messengers—and are thought to have evolved from the fusion of two CARF domains.[51] Cmr6-TIR-SAVED, thus, represents the fusion of components from three distinct bacterial immune systems. The single example of Cmr6-TIR-SAVED comes from an Alphaproteobacterial metagenomic scaffold and is encoded in a locus containing III-B Cascade components (i.e., Cmr1 through Cmr6 and a CRISPR) (Figure 10.3A).

To better understand how Cmr6-TIR-SAVED proteins are likely to be integrated into Cascade, we superimposed an Alphafold-predicted structural model of this fusion protein onto structures of Cmr6, SARM1, and Cap4. Like Cmr1, type III-B Cascade contains a single copy of Cmr6 near the 3'-end of the crRNA.[10] When our structural model of Cmr6-TIR-SAVED is superimposed on the Cmr6 subunit of III-B Cascade from *S. islandicus*, the TIR and SAVED domains jut away from the complex like a person extending an arm (Figure 10.3B). This structural comparison also revealed the presence of an N-terminal domain that was not detected during analysis of the Cmr6-TIR-SAVED primary sequence. This domain of unknown function is highly unstructured in the Alphafold-predicted structure of Cmr6-TIR-SAVED, suggesting that it may be disordered or else is an artifact of the Alphafold prediction software. To assess the TIR domain of Cmr6-TIR-SAVED, we superimposed the predicted structure of this fusion onto a structure of a eukaryotic TIR-domain protein (i.e., SARM1 from *Vitus rotundifolia*).[55] This analysis indicated Cmr6-TIR-SAVED contains a full-length TIR domain and that the residues required for $NADP^+$ cleavage are conserved in the active site of this domain (Figure 10.3B). Additionally, we confirmed that tryptophan and glutamate residues that are present in the active sites of bacterial TIR domain proteins with $NADP^+$ cleavage activity are conserved in Cmr6-TIR-SAVED (Figure 10.3C).[54] We then performed a similar comparison of Cmr6-TIR-SAVED with structures and sequences of SAVED domain proteins (i.e., Cap4 from cBASS systems) which indicated the conservation of two aromatic residues required to coordinate cyclic oligoadenylates (Figure 10.3B and D).[51] Taken together, these data suggest that none of the domains of Cmr6-TIR-SAVED has likely degenerated to become unfunctional or to catalyze an unexpected activity.

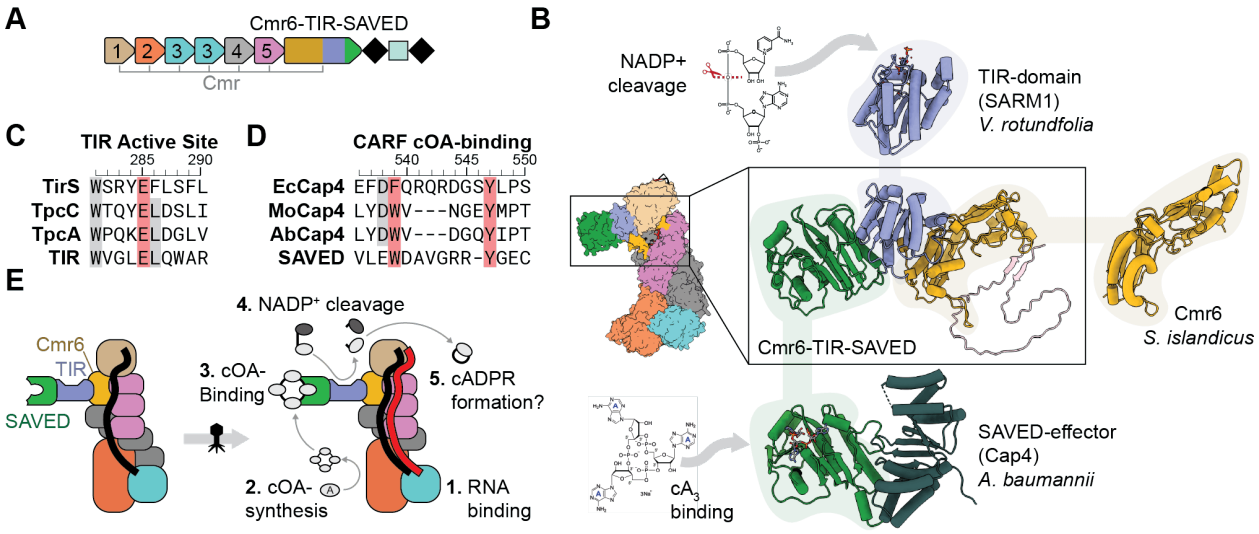


Figure 10.3. Cmr6-TIR-SAVED represents fusion of bacterial innate and adaptive immune components. (A) Genetic locus of Cmr6-TIR-SAVED contains III-B Cascade components (Cmr proteins and a CRISPR) (JGI scaffold ID: 2808606989). (B) Alphafold-predicted structure of Cmr6-TIR-SAVED superimposed on Cmr6 subunit of III-B Cascade from *S. islandicus* (left; PDB: 6SIC) suggests one copy of Cmr6-TIR-SAVED is likely present in the variant crRNA-guided surveillance complex. Cmr subunits are colored as in panel A. Predicted structure of Cmr6-TIR-SAVED also compared to TIR-domain from SARM1 (periwinkle; PDB: 6O0W), SAVED-domain from Cap4 (green; PDB: 6WAN). N-terminal Cmr6-TIR-SAVED domain of unknown function is colored pink. (C) TIR-domain active site tryptophan and glutamate residues responsible for $NADP^+$ cleavage in *Staphylococcus aureus* (TirS), *Escherichia coli* (TpcC), and *Theionarchaea archaeon* (TpcA), are conserved in Cmr6-TIR-SAVED (TIR). (D) Two aromatic residues that coordinate cOA in SAVED-domains from *Enterobacter cloacae* (EcCap4), *Moraxella osloensis* (MoCap4), and *Acinetobacter baumannii* (AbCap4) are conserved in Cmr6-TIR-SAVED (SAVED). (E) Proposed model for Cmr6-TIR-SAVED in CRISPR-mediated immunity. Binding of variant III-B Cascade to RNA target prompts cOA-formation by Cas10 subunit. cOA molecules (possibly cA_3) bind to SAVED domain, activating the $NADP^+$ cleavage activity of TIR which depletes cellular $NADP^+$ levels until cell death or dormancy is induced. NADP+ cleavage may also result in the formation of cADPR, which could activate ThsB homologs encoded elsewhere in the genome.

Based on the known functions of Cmr6-TIR-SAVED domains, we predict that this protein fusion induces cellular dormancy or apoptosis when its variant III-B surveillance complex binds to a crRNA-complementary target. Target binding by variant III-B Cascade would initiate Cas10-mediated cOA synthesis, and these second messenger molecules may then

be bound by the SAVED domain, activating the NADP+ activity of the TIR domain (Figure 10.3E). A recent analysis of nuclease effectors (i.e., Cap4) from cyclic-oligonucleotide based antiphage signaling systems (cBASS) reports that some of these effectors are activated when a cyclic-triadenylate (cA_3) molecule binds the protein's SAVED domain.[51] Among other species of cyclic-oligoadenylates, cA_3 is one of the products that type III Cascade has been shown to produce on target binding.[44,46,47] cA_3 may also be the primary ligand bound by the SAVED domain of Cmr6-TIR-SAVED, leading to activation of the TIR domain. The antiphage function of TIR domains were described in a recent pre-print from the Sorek group, in which TIR-containing proteins from Thoeris immune systems were shown to cleave NAD^+ molecules.[56] In Thoeris systems, TIR-mediated cleavage of NAD^+ results in the production cyclic ADP-Ribose (cADPR) molecules which go on to activate additional NAD^+ hydrolases (i.e., ThsB), depleting the cell of NAD^+ and arresting growth. While no ThsB homologs are encoded near the Cmr6-TIR-SAVED protein, it is tempting to speculate that the NAD^+ cleavage activity of the fusion protein's TIR domain may be sufficient to protect cells by depleting cellular NAD^+ levels. Additionally, it should be noted that Cmr6-TIR-SAVED occurs on a 26kbp metagenomic scaffold, so additional ThsB homologs may be present elsewhere in the genome of the organism from which this system originates.

Discussion

Collectively, this works reveals that Cas7 proteins from type I and type IV systems are likely to have evolved from a more recent common ancestor than Cas7 proteins from type III systems. These data support the proposed evolutionary origins of type I and IV systems and highlight the diversification and repurposing of Cas components in the emergence of novel

CRISPR-Cas subtypes.[1,2,14] The wild diversity of CRISPR systems seen today may in part be due to the arms-race nature of host-pathogen conflicts, since phage-encoded anti-CRISPR (Acr) proteins are speculated to impose strong selective pressures on CRISPR systems.[57-60] Indeed, multiple Acr proteins are known to bind Cas7 Cascade subunits (e.g., AcrIF1, AcrIF9, AcrIF14), preventing target cleavage during CRISPR-mediated interference.[3,4,61,62] Diversification of Cas7 and other Cas proteins may allow CRISPR systems to escape the select pressures imposed by Acr-mediated immune inhibition.

Our work also highlights several examples of Cas7 proteins that have been genetically fused to other Cas proteins. These fusions are particularly interesting, since they imply that multiple copies of transposition (i.e., TniQ) or HD-nuclease (i.e. Cas10 and Cas3) machinery will be present in a single crRNA-guided surveillance complex. While our bioinformatic approach is unable to resolve whether or not these variant systems are functional for CRISPR-mediated immunity or transposition, these discoveries do raise questions about how Cas protein fusions impact the mechanisms of each composite Cas domain. For example, it is unclear how HD nuclease activity would be regulated in Cas3-Cas7e and Cas10-Csm3 fusions. Unmitigated HD activity would likely be toxic to a cell, so it is kept tightly controlled in CRISPR-Cas systems. Cas3 HD activity is suppressed by Cas1 proteins in some CRISPR systems, while conformational changes in type III Cascade are thought to regulate Cas10 HD activity.[39,63,64] Perhaps Cas3-Cas7e and Cas10-Csm3 fusions rely on similar mechanisms of Cas-mediated HD suppression or else on conformational change-mediated activation of HD nuclease activity to prevent collateral autoimmunity.

The discovery of *cas7* genes that have been fused to genes from other innate immune systems (i.e., Cmr6-TIR-SAVED) and from toxin-antitoxin (TA) systems (i.e., Cmr1-MntA), shows how CRISPR systems can become diversified and gain new weapons in the fight against phages.[65] Cmr6-TIR-SAVED represents a genetic linkage between cyclic-oligonucleotide based antiphage signaling systems (cBASS) and type III CRISPR systems, which is notable since both of these systems rely on cOA-gated effectors that induce cell dormancy or apoptosis.[45,52] Thus, the marriage of these innate and adaptive immune systems in Cmr6-TIR-SAVED suggests that shared immune signaling pathways allow for the repurposing of immune components between systems.

The coupling of MntA to Cascade, via fusion to Cmr1, is reminiscent of our recent discovery that Cas10 has been fused to a TA component (i.e., AbiEii) in some type III CRISPR systems (see Chapter 9). Previous reports have detailed how TA systems contribute components to CRISPR-Cas systems—including Cas2 adaptation proteins from VapD toxins—and Cmr1-MntA appears to be a widespread (i.e., occurring in 37 open-reading from our dataset) example of this phenomenon.[2,65] While additional work is needed to determine if the HEPN toxins present in Cmr1-MntA loci participate in immune defense, it is interesting to note that type III HEPN-based effectors (e.g., Csm6 and Csx1) are also thought to have evolved from coopted TA components.[2,14] Cmr1-MntA systems could be evolutionarily related to cOA-gated type III effectors, if Cmr1 proteins lost MntA domains as HEPN RNAses from MntA systems gained CARF domains. Alternatively, Cmr1-MntA systems may represent an evolutionarily distinct example of CRISPR systems repurposing TA components for immune defense, a trend that is increasingly recognized as commonplace in the evolution of CRISPR-Cas defense systems.

Acknowledgements

B.W. is the founder of SurGene, LLC, and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof. Research in the Wiedenheft lab is supported by the National Institutes of Health (R35GM134867) and the Montana State University Agricultural Experimental Station (USDA NIFA).

References

1 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).

2 Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, doi:10.1038/s41579-019-0299-x (2019).

3 Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-57 e11, doi:10.1016/j.cell.2017.03.012 (2017).

4 Guo, T. W. *et al.* Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426 e412, doi:10.1016/j.cell.2017.09.006 (2017).

5 Huo, Y. *et al.* Cryo-EM structure of Type III-A CRISPR effector complex. *Cell Res* **28**, 1195-1197, doi:10.1038/s41422-018-0115-6 (2018).

6 Jackson, R. N. *et al.* Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science* **345**, 1473-1479, doi:10.1126/science.1256328 (2014).

7 Liu, T. Y., Liu, J. J., Aditham, A. J., Nogales, E. & Doudna, J. A. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. *Nat Commun* **10**, 3001, doi:10.1038/s41467-019-10780-2 (2019).

8 O'Brien, R. E. *et al.* Structural basis for assembly of non-canonical small subunits into type I-C Cascade. *Nat Commun* **11**, 5931, doi:10.1038/s41467-020-19785-8 (2020).

9 Rouillon, C. *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell* **52**, 124-134, doi:10.1016/j.molcel.2013.08.020 (2013).

10 Sofos, N. *et al.* Structures of the Cmr-beta Complex Reveal the Regulation of the Immunity Mechanism of Type III-B CRISPR-Cas. *Mol Cell* **79**, 741-757 e747, doi:10.1016/j.molcel.2020.07.008 (2020).

11 Zhou, Y. *et al.* Structure of a type IV CRISPR-Cas ribonucleoprotein complex. *iScience* **24**, 102201, doi:10.1016/j.isci.2021.102201 (2021).

12 Moya-Beltran, A. *et al.* Evolution of Type IV CRISPR-Cas Systems: Insights from CRISPR Loci in Integrative Conjugative Elements of Acidithiobacillia. *CRISPR J* **4**, 656-672, doi:10.1089/crispr.2021.0051 (2021).

13 Makarova, K. S. & Koonin, E. V. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol* **1311**, 47-75, doi:10.1007/978-1-4939-2687-9_4 (2015).

14 Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180087, doi:10.1098/rstb.2018.0087 (2019).

15 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

16 Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* **46**, W246-W251, doi:10.1093/nar/gky425 (2018).

17 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

18 Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-682, doi:10.1093/bioinformatics/btq003 (2010).

19 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

20 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518, doi:10.1093/nar/gki198 (2005).

21 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

22 Yu, G., Lam, T. T., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**, 3041-3043, doi:10.1093/molbev/msy194 (2018).

23 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

24 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419, doi:10.1093/nar/gkaa913 (2021).

25 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237-2243, doi:10.1016/j.jmb.2017.12.007 (2018).

26 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).

27 Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14-25, doi:10.1002/pro.3235 (2018).

28 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

29 Romero, G., von Zastrow, M. & Friedman, P. A. Role of PDZ proteins in regulating trafficking, signaling, and function of GPCRs: means, motif, and opportunity. *Adv Pharmacol* **62**, 279-314, doi:10.1016/B978-0-12-385952-5.00003-8 (2011).

30 Peters, J. E. Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond. *Mol Microbiol* **112**, 1635-1644, doi:10.1111/mmi.14383 (2019).

31 Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernandez, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. *Nature* **577**, 271-274, doi:10.1038/s41586-019-1849-0 (2020).

32 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225, doi:10.1038/s41586-019-1323-z (2019).

33 Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48-53, doi:10.1126/science.aax9181 (2019).

34 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci U S A* **114**, E7358-E7366, doi:10.1073/pnas.1709035114 (2017).

35 Rybarski, J. R. H., Kuang; Hill, Alexis M.; Wilke, Claus O. ; Finkelstein, Ilya J. Metagenomic Discovery of CRISPR-Associated Transposons. *bioRxiv*, doi:10.1101/2021.08.16.456562 (2021).

36 Wiegand, T. & Wiedenheft, B. CRISPR Surveillance Turns Transposon Taxi. *CRISPR J* **3**, 10-12, doi:10.1089/crispr.2020.29081.twi (2020).

37 Jia, N., Xie, W., de la Cruz, M. J., Eng, E. T. & Patel, D. J. Structure-function insights into the initial step of DNA integration by a CRISPR-Cas-Transposon complex. *Cell Res*, doi:10.1038/s41422-019-0272-2 (2020).

38 Wang, B., Xu, W. & Yang, H. Structural basis of a Tn7-like transposase recruitment and DNA loading to CRISPR-Cas surveillance complex. *Cell Res*, doi:10.1038/s41422-020-0274-0 (2020).

39 Rollins, M. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci U S A* **114**, E5113-E5121, doi:10.1073/pnas.1616395114 (2017).

40 Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* **21**, 771-777, doi:10.1038/nsmb.2875 (2014).

41 Rollins, M. F. *et al.* Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol Cell* **74**, 132-142 e135, doi:10.1016/j.molcel.2019.02.001 (2019).

42 Wang, X. *et al.* Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. *Nat Struct Mol Biol* **23**, 868-870, doi:10.1038/nsmb.3269 (2016).

43 van Erp, P. B. G. *et al.* Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. *ACS Chem Biol* **13**, 481-490, doi:10.1021/acschembio.7b00649 (2018).

44 Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609, doi:10.1126/science.aao0100 (2017).

45 Athukoralage, J. S. & White, M. F. Cyclic oligoadenylate signalling and regulation by ring nucleases during type III CRISPR defence. *RNA*, doi:10.1261/rna.078739.121 (2021).

46 Jia, N., Jones, R., Sukenick, G. & Patel, D. J. Second Messenger cA4 Formation within the Composite Csm1 Palm Pocket of Type III-A CRISPR-Cas Csm Complex and Its Release Path. *Mol Cell* **75**, 933-943 e936, doi:10.1016/j.molcel.2019.06.013 (2019).

47 Niewoehner, O. *et al.* Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548, doi:10.1038/nature23467 (2017).

48 Songailiene, I. *et al.* HEPN-MNT Toxin-Antitoxin System: The HEPN Ribonuclease Is Neutralized by OligoAMPylation. *Mol Cell* **80**, 955-970 e957, doi:10.1016/j.molcel.2020.11.034 (2020).

49 Yao, J. *et al.* Novel polyadenylylation-dependent neutralization mechanism of the HEPN/MNT toxin/antitoxin system. *Nucleic Acids Res* **48**, 11054-11067, doi:10.1093/nar/gkaa855 (2020).

50 Cohen, D. *et al.* Cyclic GMP-AMP signalling protects bacteria against viral infection. *Nature* **574**, 691-695, doi:10.1038/s41586-019-1605-5 (2019).

51 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* **182**, 38-49 e17, doi:10.1016/j.cell.2020.05.019 (2020).

52 Millman, A., Melamed, S., Amitai, G. & Sorek, R. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat Microbiol* **5**, 1608-1615, doi:10.1038/s41564-020-0777-y (2020).

53 Whiteley, A. T. *et al.* Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. *Nature* **567**, 194-199, doi:10.1038/s41586-019-0953-5 (2019).

54 Essuman, K. *et al.* TIR Domain Proteins Are an Ancient Family of NAD(+)-Consuming Enzymes. *Curr Biol* **28**, 421-430 e424, doi:10.1016/j.cub.2017.12.024 (2018).

55 Horsefield, S. *et al.* NAD(+) cleavage activity by animal and plant TIR domains in cell death pathways. *Science* **365**, 793-799, doi:10.1126/science.aax1911 (2019).

56 Ofir, G. *et al.* Antiviral activity of bacterial TIR domains via signaling molecules that trigger cell death. *bioRxiv*, 2021.2001.2006.425286, doi:10.1101/2021.01.06.425286 (2021).

57 Wiegand, T., Karambelkar, S., Bondy-Denomy, J. & Wiedenheft, B. Structures and Strategies of Anti-CRISPR-Mediated Immune Suppression. *Annu Rev Microbiol* **74**, 21-37, doi:10.1146/annurev-micro-020518-120107 (2020).

58 Bondy-Denomy, J. *et al.* Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136-139, doi:10.1038/nature15254 (2015).

59 Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429-432, doi:10.1038/nature11723 (2013).

60 Pawluk, A., Bondy-Denomy, J., Cheung, V. H., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. *MBio* **5**, e00896, doi:10.1128/mBio.00896-14 (2014).

61 Gabel, C., Li, Z., Zhang, H. & Chang, L. Structural basis for inhibition of the type I-F CRISPR-Cas surveillance complex by AcrIF4, AcrIF7 and AcrIF14. *Nucleic Acids Res* **49**, 584-594, doi:10.1093/nar/gkaa1199 (2021).

62 Hirschi, M. *et al.* AcrIF9 tethers non-sequence specific dsDNA to the CRISPR RNA-guided surveillance complex. *Nat Commun* **11**, 2730, doi:10.1038/s41467-020-16512-1 (2020).

63 Jung, T. Y. *et al.* Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* **23**, 782-790, doi:10.1016/j.str.2015.01.021 (2015).

64 Jia, N. *et al.* Type III-A CRISPR-Cas Csm Complexes: Assembly, Periodic RNA Cleavage, DNase Activity Regulation, and Autoimmunity. *Mol Cell* **73**, 264-277 e265, doi:10.1016/j.molcel.2018.11.007 (2019).

65 Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet* **21**, 119-131, doi:10.1038/s41576-019-0172-9 (2020).

Supplemental Tables

CasFinder HMM	CRISPR Subtype	Protein
TIGR02583	I-A	Cas7a
TIGR02589	I-C	Cas7c
TIGR01869	I-E	Cas7e
TIGR01875	I	Cas7
TIGR02585	I-B	Cas7b
TIGR03489	I-C	Cas7c
TIGR02590	I-B	Cas7b
TIGR01595	I-B	Cas7b
cmr1_IIIB_maka_4	III-B	Cmr1b
cmr1_IIIC_maka_1	III-C	Cmr1c
TIGR02580	III-B	Cmr4b
cmr4_IIIB_maka	III-B	Cmr4b
cmr6_IIIB_maka_1	III-B	Cmr6b
cmr6_IIIC_maka	III-B	Cmr6c
cmr6_IIIB_maka_2	III-B	Cmr6b
TIGR03115	IV	Csf2
csf2_IV_maka_1	IV	Csf2
csm3_IIIAD_maka_1	III-A/D	Csm3a/d
TIGR02582	III-A	Csm3a
csm3_IIID_maka_5	III-D	Csm3d
csm3_IIIAD_maka_5	III-A/D	Csm3a/d
csm3_IIID_maka_6	III-D	Csm3d
TIGR01899	III-A	Csm5a
csm5_IIIA_maka_3	III-A	Csm5b
TIGR02566	I-F	Cas7f

Supplemental Table S10.1. Profiles HMMs from CasFinder used to identify Cas7 proteins in genomic and metagenomic databases.

GENERAL CONCLUSION AND OUTLOOK

Viruses are ubiquitous genetic parasites that prey on all domains of life. Accordingly, bacteria and archaea have evolved sophisticated adaptive immune systems that protect them from these invaders. This struggle, at the interface of a genetic arms race, has produced an incredible diversity of CRISPR system variations that continues to expand as increasing numbers of genomic and metagenomic sequences become available. The applications of CRISPR-based technologies are likewise expanding rapidly, and now include wildly imaginative solutions like Cas12-based DNA-nanocage platforms for drug delivery and gene editing systems that are controlled by light.[1,2]

The diversification of CRISPR-Cas systems is fueled, in part, by the repurposing of components that are spread by mobile genetic elements (MGEs). Eugene Koonin has discussed this phenomenon, calling MGE-encoded genes ‘guns for hire’, and it is exciting to consider what other novel systems await discovery.[3] Some of the systems that seed future biotechnologies may also have yet to arisen. Indeed, work by Kim Seed’s lab that focuses on the evolution of immunity in pandemic *Vibrio cholerae*, and corresponding countermeasures adopted by its primary phage ICP1, suggests that reshuffling of immune components can reshape the landscape of the phage-host arms race multiple times in the course of a single decade.[4,5]

The use of CRISPR-Cas machinery as diagnostic tools presents an enthusing example of how bacterial and archaeal immune components can be repurposed as biotechnologies. Already, these technologies have been applied to the detection of viral, bacterial, and parasite infections, SNPs that cause genetic diseases, and miRNAs in patients with medulloblastoma.[6] Since RNA

and DNA are components of all living cells, the applications of CRISPR-based diagnostics will likely continue to expand to incorporate countless innovative uses.

Adaptation in CRISPR-Cas systems

During the course of my doctoral work, considerable progress has been made to improve our mechanistic understanding of CRISPR adaptation. *In vitro* and *in vivo* spacer acquisition assays have revealed differences in the adaptive mechanisms of distinct CRISPR systems. The I-E CRISPR system of *E. coli*, for example, requires only Cas1 and Cas2 for new sequence adaptation,[7] while Csy and Cas9 are also required for efficient acquisition in type I-F and type II systems.[8-10] While future work is needed to elucidate the mechanisms of Csy-mediated spacer integration, our work indicates that Csy may play a role in the selection of new spacer sequences and, particularly in the acquisition of PAM-proximal sequences (**Chapter 1**). We show that some sequences are preferentially acquired as new spacers in greater abundances in multiple experiments, suggesting that sequence or structural features of foreign DNA are recognized as antigenic by Cas adaptation proteins (**Chapter 1**). It should also be noted, that although the requirement for Csy and Cas9 in adaptation emerges as a parallel feature of type I-F and type II systems, recent structures of type II adaptation complexes include proteins that are not encoded in type I-F *cas* operons.[11] These findings raise additional questions about how systems that lack ancillary Cas adaptation proteins (e.g., I-F and I-E) functionally compensate for these missing modules (e.g., Cas4 and Csn2).

Our work to identify the distribution and spacing of sequence motifs in CRISPR leader elements also speaks to the parallel mechanisms of adaptation between some CRISPR subtypes. This work reveals that IHF binding sites and other motifs are found in I-C, I-F, and I-E leaders

and that the phasing (i.e., spacing) of leader motifs is critical to the integration of new spacer sequences (**Chapter 2)**. However, subtype-specific differences do emerge when one examines the pattern and phasing of I-E and I-F CRISPR leader motifs, which we use structural modeling to explain likely accommodates differences in I-E and I-F adaptation proteins (e.g. Cas1-Cas2 and Cas1-2/3, respectively) (**Chapter 2**).

SARS-CoV-2 and the COVID-19 Pandemic

The emergence of SARS-CoV-2, and the international COVID-19 pandemic, shifted the trajectory of my research efforts toward tracking the spread of this coronavirus and understanding how it evolves. This work began by monitoring levels of SARS-CoV-2 in municipal wastewater—a strategy that was widely adopted and used to monitor viral prevalence in numerous communities across the world (**Chapter 6**).[12] These data were collected early enough in the pandemic (i.e., March 2020) that limited SARS-CoV-2 genetic diversity and international travel restrictions still made it possible to identify the originating location of a SARS-CoV-2 isolate by comparing its mutational profile to other available CoV-2 genomes. Using this approach, we showed that a SARS-CoV-2 genome sequenced from wastewater in Montana was most closely related to sequences from California, Australia, and New York (**Chapter 6**). This approach at genomic epidemiology would be much more complicated today, since the genetic diversity of SARS-CoV-2 has swollen with increased spread and time since the initial emergence of the virus. Many world governments have also relaxed travel restrictions in the past year, increasing the odds that two or more genomic sequences with identical mutational profiles will appear in multiple geographically distant locales simultaneously.

When talking to friends and family in the past two years, I have compared researching SARS-CoV-2 to what it must have felt like to be an aeronautics engineer in the 1960's race to the moon. Thanks to the intense focus and collaboration of international scientists, we know more about SARS-CoV-2 today than we have about any other virus in history. Efforts to understand the contributions of every gene in the SARS-CoV-2 genome have improved our understanding of how coronaviruses replicate and how they modulate human immune systems. Our work, for example, investigating naturally occurring SARS-CoV-2 mutants with truncations in ORF7a, reveals that ORF7a is involved in dampening type I interferon responses (**Chapter 7**).

At the time of writing this dissertation chapter (November 2021), the pandemic spread of SARS-CoV-2 has unfortunately not abated, despite broad availability of numerous FDA-approved vaccines and government-implemented public health measures. In this landscape, rapid and affordable SARS-CoV-2 diagnostics are increasingly important to catch epidemic spreading events in unvaccinated communities and vaccine-breakthrough infections. Type V (i.e., Cas12) and IV (i.e., Cas13) CRISPR-based diagnostic tools are presently making their way to consumer and medical markets, and type III CRISPR-based tools will be quick to follow. These technologies are particularly powerful due to their affordability, speed, specificity, and lack of need for specialized laboratory equipment. Additionally, since crRNA guides can easily be substituted or updated, CRISPR-based diagnostics offer the potential to target novel SARS-CoV-2 variants as they emerge.

Our work shows that Csm6 and the Csm complex from *Thermus thermophilus* can be repurposed to provide visible readouts when SARS-CoV-2 RNA is present in patient samples, using a workflow that take only thirty minutes (**Chapter 8**). Three other groups reported similar

type III CRISPR-based diagnostics around the time this work was published.[13-15] While these Csm-based technologies are not currently capable of detecting clinically relevant levels of SARS-CoV-2 RNA in patient samples without a pre-amplification step—as some Cas12- and Cas13-based technologies are—additional methods to improve the sensitivities of these nascent biotechnologies are already under investigation. Indeed, our work to understand the phylogenetic and functional diversity of Cas10 (**Chapter 9**) and Cas7 (**Chapter 10**) proteins led to the identification of novel *cas* gene fusions that could be incorporated to improve the sensitivities of these technologies.

Concluding Remarks

During the course of my doctoral work, 'CRISPR' has become a term known by the general public, whereas it received blank stares from laymen when I first joined the Wiedenheft lab in the summer of 2017. These prokaryotic defense systems turned gene-editing bio-tools have already been used to cure genetic diseases and edit humans that are alive today. My work has contributed to the field of CRISPR research by revealing mechanisms of CRISPR adaptation and showing how type III systems can be repurposed to detect viral pathogens (e.g., SARS-CoV-2). Looking forward, the field of CRISPR-Cas research and the biotechnological applications of these systems have become as diverse as CRISPR systems themselves and will assuredly continue adapting to meet the future needs of humanity.

References

1 English, M. A. *et al.* Programmable CRISPR-responsive smart materials. *Science* **365**, 780-785, doi:10.1126/science.aaw5122 (2019).

2 Bubeck, F. *et al.* Engineered anti-CRISPR proteins for optogenetic control of CRISPR-Cas9. *Nat Methods* **15**, 924-927, doi:10.1038/s41592-018-0178-9 (2018).

3 Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet*, 1-13 (2019).

4 Boyd, C. M. *et al.* Bacteriophage ICP1: A Persistent Predator of Vibrio cholerae. *Annu Rev Virol* **8**, 285-304, doi:10.1146/annurev-virology-091919-072020 (2021).

5 Angermeyer, A. *et al.* Evolutionary sweeps of subviral parasites and their phage host bring unique parasite variants and disappearance of a phage CRISPR-Cas system. *bioRxiv*, 2021.2010.2007.463549, doi:10.1101/2021.10.07.463549 (2021).

6 Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nat Biomed Eng* **5**, 643-656, doi:10.1038/s41551-021-00760-7 (2021).

7 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res* **40**, 5569-5576, doi:10.1093/nar/gks216 (2012).

8 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202, doi:10.1038/nature14245 (2015).

9 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).

10 Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-361, doi:10.1101/gad.257550.114 (2015).

11 Wilkinson, M. *et al.* Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. *Mol Cell* **75**, 90-101 e105, doi:10.1016/j.molcel.2019.04.020 (2019).

12 Kitajima, M. *et al.* SARS-CoV-2 in wastewater: State of the knowledge and research needs. *Sci Total Environ* **739**, 139076, doi:10.1016/j.scitotenv.2020.139076 (2020).

13 Sridhara, S., Goswami, H. N., Whyms, C., Dennis, J. H. & Li, H. Virus detection via programmable Type III-A CRISPR-Cas systems. *Nat Commun* **12**, 5653, doi:10.1038/s41467-021-25977-7 (2021).

14 Steens, J. A. *et al.* SCOPE enables type III CRISPR-Cas diagnostics using flexible targeting and stringent CARF ribonuclease activation. *Nat Commun* **12**, 5033, doi:10.1038/s41467-021-25337-5 (2021).

15 Grüschow, S., Adamson, C. S. & White, M. F. Specificity and sensitivity of an RNA targeting type III CRISPR complex coupled with a NucC endonuclease effector. *bioRxiv*, 2021.2009.2013.460032, doi:10.1101/2021.09.13.460032 (2021).

REFERENCES CITED

Addetia, A. et al. Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. J Clin Virol 129, 104523, doi:10.1016/j.jcv.2020.104523 (2020).

Aeling, K. A. et al. Indirect Recognition in Sequence-specific DNA Binding by Escherichia coli Integration Host Factor. Journal of Biological Chemistry 281, 39236-39248, doi:10.1074/jbc.m606363200 (2006).

Ahmed, W. et al. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. Sci Total Environ 728, 138764-138764, doi:10.1016/j.scitotenv.2020.138764 (2020).

Almendros, C., Guzman, N. M., Garcia-Martinez, J. & Mojica, F. J. Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. Nat Microbiol 1, 16081, doi:10.1038/nmicrobiol.2016.81 (2016).

Amitai, G. & Sorek, R. CRISPR–Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol 14, 67 (2016).

Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513, 569 (2014).

Andolfo, I. et al. Common variants at 21q22.3 locus influence MX1 and TMPRSS2 gene expression and susceptibility to severe COVID-19. iScience 24, 102322, doi:10.1016/j.isci.2021.102322 (2021).

Andrew Rambaut, N. L., Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz, on behalf of COVID-19 Genomics Consortium UK (CoG-UK). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, 2020).

Angermeyer, A. et al. Evolutionary sweeps of subviral parasites and their phage host bring unique parasite variants and disappearance of a phage CRISPR-Cas system. bioRxiv, 2021.2010.2007.463549, doi:10.1101/2021.10.07.463549 (2021).

Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science 369, 1210-1220, doi:10.1126/science.abc6261 (2020).

Aschenbrenner, S. et al. Coupling Cas9 to artificial inhibitory domains enhances CRISPR-Cas9 target specificity. Sci Adv 6, eaay0187, doi:10.1126/sciadv.aay0187 (2020).

Athukoralage, J. S. et al. The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling. Elife 9, doi:10.7554/eLife.55852 (2020).

Athukoralage, J. S. et al. An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. Nature 577, 572-575, doi:10.1038/s41586-019-1909-5 (2020).

Athukoralage, J. S. & White, M. F. Cyclic oligoadenylate signalling and regulation by ring nucleases during type III CRISPR defence. RNA, doi:10.1261/rna.078739.121 (2021).

Avanzato, V. A. et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. Cell 183, 1901-1912 e1909, doi:10.1016/j.cell.2020.10.049 (2020).

Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research 37, W202-W208, doi:10.1093/nar/gkp335 (2009).

Banerjee, A. et al. Experimental and natural evidence of SARS-CoV-2-infection-induced activation of type I interferon responses. iScience 24, 102477, doi:10.1016/j.isci.2021.102477 (2021).

Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709-1712 (2007).

Barrangou, R. et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315, 1709-1712, doi:10.1126/science.1138140 (2007).

Basgall, E. M. et al. Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and AcrIIA4 in Saccharomyces cerevisiae. Microbiology 164, 464-474, doi:10.1099/mic.0.000635 (2018).

Bedford, T. et al. Cryptic transmission of SARS-CoV-2 in Washington state. Science 370, 571-575, doi:10.1126/science.abc0523 (2020).

Beloglazova, N. et al. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. EMBO J 30, 4616-4627, doi:10.1038/emboj.2011.377 (2011).

Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. Nucleic acids research 48, 748-760, doi:10.1093/nar/gkz1091 (2020).

Bhoobalan-Chitty, Y., Johansen, T. B., Di Cianni, N. & Peng, X. Inhibition of Type III CRISPR-Cas Immunity by an Archaeal Virus-Encoded Anti-CRISPR Protein. Cell 179, 448-458. e411 (2019).

Biggerstaff, M. et al. Influenza-like illness, the time to seek healthcare, and influenza antiviral receipt during the 2010-2011 influenza season-United States. J Infect Dis 210, 535-544, doi:10.1093/infdis/jiu224 (2014).

Biosciences, M. SARS-CoV-2 RNA DETECTR Reagent Kit. (https://www.fda.gov/media/141765/download, 2020).

Biosciences, S. Instructions for use: SherlockTM CRISPR SARS-CoV-2 kit. (https://www.fda.gov/media/137746/download, 2020).

Birkholz, N., Fagerlund, R. D., Smith, L. M., Jackson, S. A. & Fineran, P. C. The autoregulator Aca2 mediates anti-CRISPR repression. Nucleic Acids Res 47, 9658-9665, doi:10.1093/nar/gkz721 (2019).

Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 17, 356-356, doi:10.1186/s12864-016-2627-0 (2016).

Blanco-Melo, D. et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell 181, 1036-1045 e1039, doi:10.1016/j.cell.2020.04.026 (2020).

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151, 2551-2561, doi:10.1099/mic.0.28048-0 (2005).

Bondy-Denomy, J. et al. A unified resource for tracking anti-CRISPR names. The CRISPR journal 1, 304-305 (2018).

Bondy-Denomy, J. et al. Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature 526, 136-139, doi:10.1038/nature15254 (2015).

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature 493, 429-432, doi:10.1038/nature11723 (2013).

Borges, A. L., Davidson, A. R. & Bondy-Denomy, J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. Annu Rev Virol 4, 37-59, doi:10.1146/annurev-virology-101416-041616 (2017).

Borges, A. L. et al. Bacteriophage Cooperation Suppresses CRISPR-Cas3 and Cas9 Immunity. Cell 174, 917-925.e910, doi:10.1016/j.cell.2018.06.013 (2018).

Boyd, C. M. et al. Bacteriophage ICP1: A Persistent Predator of Vibrio cholerae. Annu Rev Virol 8, 285-304, doi:10.1146/annurev-virology-091919-072020 (2021).

Broughton, J. P. et al. CRISPR-Cas12-based detection of SARS-CoV-2. Nat Biotechnol 38, 870-874, doi:10.1038/s41587-020-0513-4 (2020).

Bryson, A. L. et al. Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. MBio 6, e00648, doi:10.1128/mBio.00648-15 (2015).

Bubeck, F. et al. Engineered anti-CRISPR proteins for optogenetic control of CRISPR-Cas9. Nat Methods 15, 924-927, doi:10.1038/s41592-018-0178-9 (2018).

Budhathoki, J. B. et al. Real-time observation of CRISPR spacer acquisition by Cas1-Cas2 integrase. Nat Struct Mol Biol 27, 489-499, doi:10.1038/s41594-020-0415-7 (2020).

Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. J Bacteriol 194, 5728-5738, doi:10.1128/JB.01184-12 (2012).

Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421-421, doi:10.1186/1471-2105-10-421 (2009).

Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

Carte, J. et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus. Molecular microbiology 93, 98-112, doi:10.1111/mmi.12644 (2014).

Casteel, D. E. et al. A DNA polymerase-{alpha}{middle dot}primase cofactor with homology to replication protein A-32 regulates DNA replication in mammalian cells. J Biol Chem 284, 5807-5818, doi:10.1074/jbc.M807593200 (2009).

Castillo, A. E. et al. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. J Med Virol 92, 1562-1566, doi:10.1002/jmv.25797 (2020).

Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS microbiology reviews 39, 428-441 (2015).

Charpentier, E., van der Oost, J. & White, M. F. in CRISPR-Cas Systems 115-144 (Springer, 2013).

Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science 360, 436-439, doi:10.1126/science.aar6245 (2018).

Cheng, K. C. & Smith, G. R. Recombinational hotspot activity of Chi-like sequences. J Mol Biol 180, 371-377, doi:10.1016/s0022-2836(84)80009-1 (1984).

Cheng, K. C. & Smith, G. R. Cutting of chi-like sequences by the RecBCD enzyme of Escherichia coli. J Mol Biol 194, 747-750, doi:10.1016/0022-2836(87)90252-x (1987).

Chinese, S. M. E. C. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 303, 1666-1669, doi:10.1126/science.1092002 (2004).

Choi, B. et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. N Engl J Med 383, 2291-2293, doi:10.1056/NEJMc2031364 (2020).

Chowdhury, S. et al. Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell 169, 47-57 e11, doi:10.1016/j.cell.2017.03.012 (2017).

Cocozaki, A. I. et al. Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. Structure 20, 545-553, doi:10.1016/j.str.2012.01.018 (2012).

Cohen, D. et al. Cyclic GMP-AMP signalling protects bacteria against viral infection. Nature 574, 691-695, doi:10.1038/s41586-019-1605-5 (2019).

Connolly, M. et al. Static Kinks or Flexible Hinges: Multiple Conformations of Bent DNA Bound to Integration Host Factor Revealed by Fluorescence Lifetime Measurements. J Phys Chem B 122, 11519-11534, doi:10.1021/acs.jpcb.8b07405 (2018).

Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 46, W246-W251, doi:10.1093/nar/gky425 (2018).

Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res 14, 1188-1190, doi:10.1101/gr.849004 (2004).

Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. Science advances 3, eaao0027 (2017).

Dao Thi, V. L. et al. A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. Sci Transl Med 12, doi:10.1126/scitranslmed.abc7075 (2020).

Datsenko, K. A. & Wanner, B. L. One-Step Inactivation of Chromosomal Genes in Escherichia coli K-12 Using PCR Products. Proceedings of the National Academy of Sciences of the United States of America 97, 6640-6645 (2000).

Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372, doi:10.1126/science.abg3055 (2021).

Deng, X. et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science 369, 582-587, doi:10.1126/science.abb9263 (2020).

Deveau, H. et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol 190, 1390-1400, doi:10.1128/JB.01412-07 (2008).

Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev 72, 642-671, Table of Contents, doi:10.1128/MMBR.00020-08 (2008).

Dolan, A. E. et al. Introducing a Spectrum of Long-Range Genomic Deletions in Human Embryonic Stem Cells Using Type I CRISPR-Cas. Mol Cell 74, 936-950 e935, doi:10.1016/j.molcel.2019.03.014 (2019).

Dolgin, E. The kill-switch for CRISPR that could make gene-editing safer. Nature 577, 308-310, doi:10.1038/d41586-020-00053-0 (2020).

Dong et al. Structural basis of CRISPR-SpyCas9 inhibition by an anti-CRISPR protein. Nature 546, 436-439, doi:10.1038/nature22377 (2017).

Dong, L. et al. An anti-CRISPR protein disables type V Cas12a by acetylation. Nat Struct Mol Biol 26, 308-314, doi:10.1038/s41594-019-0206-1 (2019).

Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. Nucleic Acids Res 42, 4590-4605, doi:10.1093/nar/gkt1419 (2014).

Eaglesham, J. B., Pan, Y., Kupper, T. S. & Kranzusch, P. J. Viral and metazoan poxins are cGAMP-specific nucleases that restrict cGAS-STING signalling. Nature 566, 259-263, doi:10.1038/s41586-019-0928-6 (2019).

East-Seletsky, A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature 538, 270-273, doi:10.1038/nature19802 (2016).

Eckweiler, D., Dudek, C.-A., Hartlich, J., Brötje, D. & Jahn, D. PRODORIC2: the bacterial gene regulation database in 2018. Nucleic acids research 46, D320-D326, doi:10.1093/nar/gkx1091 (2018).

Eddy, S. R. Accelerated Profile HMM Searches. PLoS Comput Biol 7, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

Edgar, R. et al. Bacteriophage infection is targeted to cellular poles. Mol Microbiol 68, 1107-1116, doi:10.1111/j.1365-2958.2008.06205.x (2008).

Edgar, R. C. in Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. (IEEE).

Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1, 33-46, doi:10.1002/gch2.1018 (2017).

Elde, N. C. et al. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150, 831-841, doi:10.1016/j.cell.2012.05.049 (2012).

Elmore, J. R. et al. Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. Genes Dev 30, 447-459, doi:10.1101/gad.272153.115 (2016).

Emeny, J. M. & Morgan, M. J. Regulation of the interferon system: evidence that Vero cells have a genetic defect in interferon production. J Gen Virol 43, 247-252, doi:10.1099/0022-1317-43-1-247 (1979).

English, M. A. et al. Programmable CRISPR-responsive smart materials. Science 365, 780-785, doi:10.1126/science.aaw5122 (2019).

Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. Mol Microbiol 91, 900-917, doi:10.1111/mmi.12503 (2014).

Essuman, K. et al. TIR Domain Proteins Are an Ancient Family of NAD(+)-Consuming Enzymes. Curr Biol 28, 421-430 e424, doi:10.1016/j.cub.2017.12.024 (2018).

Estrella, M. A., Kuo, F. T. & Bailey, S. RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. Genes Dev 30, 460-470, doi:10.1101/gad.273722.115 (2016).

Fagerlund, R. D. et al. Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences of the United States of America 114, E5122-E5128, doi:10.1073/pnas.1618421114 (2017).

Fauver, J. R. et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell 181, 990-996.e995, doi:10.1016/j.cell.2020.04.021 (2020).

Fozouni, P. et al. Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. Cell 184, 323-333 e329, doi:10.1016/j.cell.2020.12.001 (2021).

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

Fuchsbauer, O. et al. Cas9 Allosteric Inhibition by the Anti-CRISPR Protein AcrIIA6. Mol Cell, doi:10.1016/j.molcel.2019.09.012 (2019).

Gabel, C., Li, Z., Zhang, H. & Chang, L. Structural basis for inhibition of the type I-F CRISPR-Cas surveillance complex by AcrIF4, AcrIF7 and AcrIF14. Nucleic Acids Res 49, 584-594, doi:10.1093/nar/gkaa1199 (2021).

Gamage, A. M. et al. Infection of human Nasal Epithelial Cells with SARS-CoV-2 and a 382-nt deletion isolate lacking ORF8 reveals similar viral kinetics and host transcriptional profiles. PLoS Pathog 16, e1009130, doi:10.1371/journal.ppat.1009130 (2020).

Gao, X., Sun, B. & Guan, Y. Pullulan reduces the non-specific amplification of loop-mediated isothermal amplification (LAMP). Anal Bioanal Chem 411, 1211-1218, doi:10.1007/s00216-018-1552-2 (2019).

Garg, S. et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. MMWR Morb Mortal Wkly Rep 69, 458-464, doi:10.15585/mmwr.mm6915e3 (2020).

Garrett, S. et al. Primed CRISPR DNA uptake in Pyrococcus furiosus. Nucleic Acids Res 48, 6120-6135, doi:10.1093/nar/gkaa381 (2020).

Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci 27, 14-25, doi:10.1002/pro.3235 (2018).

Gonzalez-Reiche, A. S. et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369, 297-301, doi:10.1126/science.abc1917 (2020).

Gootenberg, J. S. et al. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science 360, 439-444, doi:10.1126/science.aaq0179 (2018).

Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science 356, 438-442, doi:10.1126/science.aam9321 (2017).

Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics 8, 312-312, doi:10.1186/1471-2105-8-312 (2007).

Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol 20, 8, doi:10.1186/s13059-018-1618-7 (2019).

Grüschow, S., Adamson, C. S. & White, M. F. Specificity and sensitivity of an RNA targeting type III CRISPR complex coupled with a NucC endonuclease effector. bioRxiv, 2021.2009.2013.460032, doi:10.1101/2021.09.13.460032 (2021).

Gudbergsdottir, S. et al. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Molecular microbiology 79, 35-49, doi:10.1111/j.1365-2958.2010.07452.x (2011).

Guo, T. W. et al. Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. Cell 171, 414-426 e412, doi:10.1016/j.cell.2017.09.006 (2017).

Hadjadj, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science 369, 718-724, doi:10.1126/science.abc6027 (2020).

Hale, C. R. et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139, 945-956, doi:10.1016/j.cell.2009.07.040 (2009).

Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernández, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. Nature 577, 271-274, doi:10.1038/s41586-019-1849-0 (2019).

Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernandez, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. Nature 577, 271-274, doi:10.1038/s41586-019-1849-0 (2020).

Harcourt, J. et al. Severe Acute Respiratory Syndrome Coronavirus 2 from Patient with Coronavirus Disease, United States. Emerg Infect Dis 26, 1266-1273, doi:10.3201/eid2606.200516 (2020).

Hardinge, P. & Murray, J. A. H. Reduced False Positives and Improved Reporting of Loop-Mediated Isothermal Amplification using Quenched Fluorescent Primers. Sci Rep 9, 7400, doi:10.1038/s41598-019-43817-z (2019).

Hart, O. E. & Halden, R. U. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. Sci Total Environ 730, 138875-138875, doi:10.1016/j.scitotenv.2020.138875 (2020).

Hatoum-Aslan, A., Maniv, I., Samai, P. & Marraffini, L. A. Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. J Bacteriol 196, 310-317, doi:10.1128/JB.01130-13 (2014).

He, F. et al. Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype ID immunity. Nat Microbiol 3, 461 (2018).

Heler, R. et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature 519, 199-202, doi:10.1038/nature14245 (2015).

Heussler, G. E., Miller, J. L., Price, C. E., Collins, A. J. & O'Toole, G. A. Requirements for Pseudomonas aeruginosa Type I-F CRISPR-Cas Adaptation Determined Using a Biofilm Enrichment Assay. J Bacteriol 198, 3080-3090, doi:10.1128/JB.00458-16 (2016).

Hidalgo-Cantabrana, C., Goh, Y. J., Pan, M., Sanozky-Dawes, R. & Barrangou, R. Genome editing using the endogenous type I CRISPR-Cas system in Lactobacillus crispatus. Proc Natl Acad Sci U S A 116, 15774-15783, doi:10.1073/pnas.1905421116 (2019).

Hille, F. et al. The biology of CRISPR-Cas: backward and forward. Cell 172, 1239-1259 (2018).

Hindson, J. COVID-19: faecal-oral transmission? Nat Rev Gastroenterol Hepatol 17, 259-259, doi:10.1038/s41575-020-0295-7 (2020).

Hirschi, M. et al. AcrIF9 tethers non-sequence specific dsDNA to the CRISPR RNA-guided surveillance complex. Nat Commun 11, 2730, doi:10.1038/s41467-020-16512-1 (2020).

Hoffmann, M. et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181, 271-280 e278, doi:10.1016/j.cell.2020.02.052 (2020).

Holland, L. A. et al. An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). J Virol 94, doi:10.1128/JVI.00711-20 (2020).

Holshue, M. L. et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med 382, 929-936, doi:10.1056/NEJMoa2001191 (2020).

Horsefield, S. et al. NAD(+) cleavage activity by animal and plant TIR domains in cell death pathways. Science 365, 793-799, doi:10.1126/science.aax1911 (2019).

Hou, Y. J. et al. SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. Cell 182, 429-446 e414, doi:10.1016/j.cell.2020.05.042 (2020).

Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680-682, doi:10.1093/bioinformatics/btq003 (2010).

Huo, Y. et al. Cryo-EM structure of Type III-A CRISPR effector complex. Cell Res 28, 1195-1197, doi:10.1038/s41422-018-0115-6 (2018).

Huo, Y. et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol 21, 771-777, doi:10.1038/nsmb.2875 (2014).

Hurford, A. & Day, T. Immune evasion and the evolution of molecular mimicry in parasites. Evolution 67, 2889-2904, doi:10.1111/evo.12171 (2013).

Hwang, S. & Maxwell, K. L. Meet the Anti-CRISPRs: Widespread Protein Inhibitors of CRISPR-Cas Systems. CRISPR J 2, 23-30, doi:10.1089/crispr.2018.0052 (2019).

Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119, doi:10.1186/1471-2105-11-119 (2010).

Hynes, A. P. et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun 9, 2919, doi:10.1038/s41467-018-05092-w (2018).

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol 169, 5429-5433, doi:10.1128/jb.169.12.5429-5433.1987 (1987).

Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS One 15, e0239403-e0239403, doi:10.1371/journal.pone.0239403 (2020).

Jackson, R. N. et al. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science 345, 1473-1479, doi:10.1126/science.1256328 (2014).

Jackson, R. N., van Erp, P. B., Sternberg, S. H. & Wiedenheft, B. Conformational regulation of CRISPR-associated nucleases. Current opinion in microbiology 37, 110-119 (2017).

Jackson, S. A., Birkholz, N., Malone, L. M. & Fineran, P. C. Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. Cell Host Microbe 25, 250-260 e254, doi:10.1016/j.chom.2018.12.014 (2019).

Jackson, S. A. et al. CRISPR-Cas: Adapting to change. Science 356, doi:10.1126/science.aal5056 (2017).

Jackson, S. A. et al. CRISPR-Cas: Adapting to change. Science 356, doi:10.1126/science.aal5056 (2017).

Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology 3, 318-356, doi:10.1016/s0022-2836(61)80072-7 (1961).

Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338-345, doi:10.1038/nbt.4060 (2018).

Jansen, R., Embden, J. D. A. v., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology 43, 1565-1575, doi:10.1046/j.1365-2958.2002.02839.x (2002).

Jassal, B. et al. The reactome pathway knowledgebase. Nucleic Acids Res 48, D498-D503, doi:10.1093/nar/gkz1031 (2020).

Jia, N., Jones, R., Sukenick, G. & Patel, D. J. Second Messenger cA4 Formation within the Composite Csm1 Palm Pocket of Type III-A CRISPR-Cas Csm Complex and Its Release Path. Mol Cell 75, 933-943 e936, doi:10.1016/j.molcel.2019.06.013 (2019).

Jia, N., Jones, R., Yang, G., Ouerfelli, O. & Patel, D. J. CRISPR-Cas III-A Csm6 CARF Domain Is a Ring Nuclease Triggering Stepwise cA4 Cleavage with ApA>p Formation Terminating RNase Activity. Mol Cell 75, 944-956 e946, doi:10.1016/j.molcel.2019.06.014 (2019).

Jia, N. et al. Type III-A CRISPR-Cas Csm Complexes: Assembly, Periodic RNA Cleavage, DNase Activity Regulation, and Autoimmunity. Mol Cell 73, 264-277 e265, doi:10.1016/j.molcel.2018.11.007 (2019).

Jia, N., Xie, W., de la Cruz, M. J., Eng, E. T. & Patel, D. J. Structure-function insights into the initial step of DNA integration by a CRISPR-Cas-Transposon complex. Cell Res, doi:10.1038/s41422-019-0272-2 (2020).

Jiang, F. et al. Temperature-responsive competitive inhibition of CRISPR-Cas9. Mol Cell 73, 601-610. e605 (2019).

Jiang, H. W. et al. SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. Cell Mol Immunol 17, 998-1000, doi:10.1038/s41423-020-0514-8 (2020).

Jones, D. L. et al. Kinetics of dCas9 target search in Escherichia coli. Science 357, 1420-1424 (2017).

Joung, J. et al. Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing. N Engl J Med 383, 1492-1494, doi:10.1056/NEJMc2026172 (2020).

Joyeux, M. Preferential Localization of the Bacterial Nucleoid. Microorganisms 7, doi:10.3390/microorganisms7070204 (2019).

Jung, T. Y. et al. Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. Structure 23, 782-790, doi:10.1016/j.str.2015.01.021 (2015).

Ka, D., An, S. Y., Suh, J.-Y. & Bae, E. Crystal structure of an anti-CRISPR protein, AcrIIA1. Nucleic Acids Res 46, 485-492 (2017).

Kamashev, D. et al. Comparison of histone-like HU protein DNA-binding properties and HU/IHF protein sequence alignment. PLoS One 12, e0188037-e0188037, doi:10.1371/journal.pone.0188037 (2017).

Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. Nat Biomed Eng 5, 643-656, doi:10.1038/s41551-021-00760-7 (2021).

Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33, 511-518, doi:10.1093/nar/gki198 (2005).

Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780, doi:10.1093/molbev/mst010 (2013).

Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science 357, 605-609, doi:10.1126/science.aao0100 (2017).

Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, C. & Siksnys, V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Mol Cell 62, 295-306, doi:10.1016/j.molcel.2016.03.024 (2016).

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10, 845-858, doi:10.1038/nprot.2015.053 (2015).

Kieper, S. N., Almendros, C. & Brouns, S. J. J. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. FEMS Microbiol Lett 366, fnz129, doi:10.1093/femsle/fnz129 (2019).

Kim, D. et al. The Architecture of SARS-CoV-2 Transcriptome. Cell 181, 914-921 e910, doi:10.1016/j.cell.2020.04.011 (2020).

Kim, J. G., Garrett, S., Wei, Y., Graveley, B. R. & Terns, M. P. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. Nucleic acids research 47, 8632-8648, doi:10.1093/nar/gkz677 (2019).

Kim, S. et al. Selective loading and processing of prespacers for precise CRISPR adaptation. Nature 579, 141-145, doi:10.1038/s41586-020-2018-1 (2020).

Kim, Y. et al. Anti-CRISPR AcrIIC3 discriminates between Cas9 orthologs via targeting the variable surface of the HNH nuclease domain. The FEBS journal (2019).

Kitajima, M. et al. SARS-CoV-2 in wastewater: State of the knowledge and research needs. Sci Total Environ 739, 139076, doi:10.1016/j.scitotenv.2020.139076 (2020).

Klein, S. et al. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. Nat Commun 11, 5885, doi:10.1038/s41467-020-19619-7 (2020).

Klompe, S. E. & Sternberg, S. H. Harnessing “A Billion Years of Experimentation”: the ongoing exploration and exploitation of CRISPR–Cas immune systems. The CRISPR journal 1, 141-158 (2018).

"Klompe, S. E. & Sternberg, S. H. Harnessing ""A Billion Years of Experimentation"": The Ongoing Exploration and Exploitation of CRISPR-Cas Immune Systems. CRISPR J 1, 141-158, doi:10.1089/crispr.2018.0012 (2018)."

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. Nature 571, 219-225, doi:10.1038/s41586-019-1323-z (2019).

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. Nature 571, 219-225, doi:10.1038/s41586-019-1323-z (2019).

Knott, G. J. et al. Structural basis for AcrVA4 inhibition of specific CRISPR-Cas12a. Elife 8, doi:10.7554/eLife.49110 (2019).

Knott, G. J. et al. Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. Nat Struct Mol Biol 26, 315-321, doi:10.1038/s41594-019-0208-z (2019).

Kobryn, K., Naigamwalla, D. Z. & Chaconas, G. Site-specific DNA binding and bending by the Borrelia burgdorferi Hbb protein. Molecular Microbiology 37, 145-155, doi:10.1046/j.1365-2958.2000.01981.x (2000).

Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci 374, 20180087, doi:10.1098/rstb.2018.0087 (2019).

Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet, 1-13 (2019).

Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet 21, 119-131, doi:10.1038/s41576-019-0172-9 (2020).

Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182, 812-827.e819, doi:10.1016/j.cell.2020.06.043 (2020).

Kotenko, S. V. & Durbin, J. E. Contribution of type III interferons to antiviral immunity: location, location, location. J Biol Chem 292, 7295-7303, doi:10.1074/jbc.R117.777102 (2017).

Kumble, K. D., Iversen, P. L. & Vishwanatha, J. K. The role of primer recognition proteins in DNA replication: inhibition of cellular proliferation by antisense oligodeoxyribonucleotides. J Cell Sci 101 (Pt 1), 35-41 (1992).

Kupczok, A., Landan, G. & Dagan, T. The Contribution of Genetic Recombination to CRISPR Array Evolution. Genome Biol Evol 7, 1925-1939, doi:10.1093/gbe/evv113 (2015).

La Rosa, G. et al. First detection of SARS-CoV-2 in untreated wastewaters in Italy. Sci Total Environ 736, 139652-139652, doi:10.1016/j.scitotenv.2020.139652 (2020).

La Scola, B. et al. Viral RNA load as determined by cell culture as a management tool for discharge of SARS-CoV-2 patients from infectious disease wards. Eur J Clin Microbiol Infect Dis 39, 1059-1061, doi:10.1007/s10096-020-03913-9 (2020).

Landsberger, M. et al. Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell 174, 908-916.e912, doi:10.1016/j.cell.2018.05.058 (2018).

Larremore, D. B. et al. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. Sci Adv 7, doi:10.1126/sciadv.abd5393 (2021).

Lau, R. K. et al. Structure and Mechanism of a Cyclic Trinucleotide-Activated Bacterial Endonuclease Mediating Bacteriophage Immunity. Mol Cell 77, 723-733 e726, doi:10.1016/j.molcel.2019.12.010 (2020).

Lauer, S. A. et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med 172, 577-582, doi:10.7326/M20-0504 (2020).

Lazer, D. S., M.; Perlis, R.H.; Ognyanova, K.; Baum, M.A.; Quintana, A.; Druckman, J.; Volpe, J.D.; Chwe, H.; Simonson, M. State of the Nation: A 50-State COVID-19 Survey: Report #8: Failing the Test: Waiting Times for COVID Diagnostic Tests Across the U.S., (World Health Organization, 2020).

Lee, D. H. & Schleif, R. F. In vivo DNA loops in araCBAD: size limits and helical repeat. Proceedings of the National Academy of Sciences of the United States of America 86, 476-480, doi:10.1073/pnas.86.2.476 (1989).

Lei, X. et al. Activation and evasion of type I interferon responses by SARS-CoV-2. Nat Commun 11, 3810, doi:10.1038/s41467-020-17665-9 (2020).

Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, doi:10.1126/science.abe3261 (2021).

Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. Nature Reviews Genetics 15, 453-468, doi:10.1038/nrg3684 (2014).

Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520, 505-510, doi:10.1038/nature14302 (2015).

Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

Li, M. et al. Toxin-antitoxin RNA pairs safeguard CRISPR-Cas systems. Science 372, doi:10.1126/science.abe5601 (2021).

Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res 42, 2483-2492, doi:10.1093/nar/gkt1154 (2014).

Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 382, 1199-1207, doi:10.1056/NEJMoa2001316 (2020).

Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

Li, Z., Zhang, H., Xiao, R. & Chang, L. Cryo-EM structure of a type I-F CRISPR RNA guided surveillance complex bound to transposition protein TniQ. Cell Res, doi:10.1038/s41422-019-0268-y (2020).

Lilly, J. & Camps, M. Mechanisms of Theta Plasmid Replication. Microbiol Spectr 3, PLAS-0029-2014, doi:10.1128/microbiolspec.PLAS-0029-2014 (2015).

Lin, J. et al. DNA targeting by subtype I-D CRISPR-Cas shows type I and type III features. Nucleic Acids Res 48, 10470-10478, doi:10.1093/nar/gkaa749 (2020).

Liu, L., Yin, M., Wang, M. & Wang, Y. Phage AcrIIA2 DNA Mimicry: Structural Basis of the CRISPR and Anti-CRISPR Arms Race. Mol Cell 73, 611-620 e613, doi:10.1016/j.molcel.2018.11.011 (2019).

Liu, T. et al. Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res 45, 8978-8992, doi:10.1093/nar/gkx612 (2017).

Liu, T. Y., Iavarone, A. T. & Doudna, J. A. RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. PLoS One 12, e0170552, doi:10.1371/journal.pone.0170552 (2017).

Liu, T. Y., Liu, J. J., Aditham, A. J., Nogales, E. & Doudna, J. A. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. Nat Commun 10, 3001, doi:10.1038/s41467-019-10780-2 (2019).

Lodder, W. & de Roda Husman, A. M. SARS-CoV-2 in wastewater: potential health risk, but also data source. Lancet Gastroenterol Hepatol 5, 533-534, doi:10.1016/S2468-1253(20)30087-X (2020).

Loman, N. J. R., W.; Rambaut, A. NCoV-2019 Novel Coronavirus Bioinformatics Protocol. (Artic Network, 2020).

Lopez-Alvarez, D., Parra, B. & Cuellar, W. J. Genome Sequence of SARS-CoV-2 Isolate Cali-01, from Colombia, Obtained Using Oxford Nanopore MinION Sequencing. Microbiol Resour Announc 9, e00573-00520, doi:10.1128/MRA.00573-20 (2020).

Loveday, E. K. et al. Effect of Inactivation Methods on SARS-CoV-2 Virion Protein and Structure. Viruses 13, doi:10.3390/v13040562 (2021).

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 182, 38-49 e17, doi:10.1016/j.cell.2020.05.019 (2020).

Lucas, C. et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature 584, 463-469, doi:10.1038/s41586-020-2588-y (2020).

Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biology direct 6, 38 (2011).

Makarova, K. S. & Koonin, E. V. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol 1311, 47-75, doi:10.1007/978-1-4939-2687-9_4 (2015).

Makarova, K. S. et al. Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res 48, 8828-8847, doi:10.1093/nar/gkaa635 (2020).

Makarova, K. S. et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13, 722-736, doi:10.1038/nrmicro3569 (2015).

Makarova, K. S. et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13, 722-736, doi:10.1038/nrmicro3569 (2015).

Makarova, K. S. et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol, doi:10.1038/s41579-019-0299-x (2019).

Mallapaty, S. How sewage could reveal true scale of coronavirus outbreak. Nature 580, 176-177, doi:10.1038/d41586-020-00973-x (2020).

Manav, M. C. et al. Structural basis for inhibition of an archaeal CRISPR-Cas type I-D large subunit by an anti-CRISPR protein. Nat Commun 11, 5993, doi:10.1038/s41467-020-19847-x (2020).

Mantlo, E., Bukreyeva, N., Maruyama, J., Paessler, S. & Huang, C. Antiviral activities of type I interferons to SARS-CoV-2 infection. Antiviral Res 179, 104811, doi:10.1016/j.antiviral.2020.104811 (2020).

Marino, N. D., Pinilla-Redondo, R., Csorgo, B. & Bondy-Denomy, J. Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. Nat Methods, doi:10.1038/s41592-020-0771-6 (2020).

McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Molecular cell 64, 616-623, doi:10.1016/j.molcel.2016.08.038 (2016).

McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat Rev Microbiol 17, 7-12 (2019).

McKenzie, R. E., Almendros, C., Vink, J. N. A. & Brouns, S. J. J. Using CAPTURE to detect spacer acquisition in native CRISPR arrays. Nat Protoc 14, 976-990, doi:10.1038/s41596-018-0123-5 (2019).

McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in infectious disease. Nat Rev Immunol 15, 87-103, doi:10.1038/nri3787 (2015).

Meagher, R. J., Priye, A., Light, Y. K., Huang, C. & Wang, E. Impact of primer dimers and self-amplifying hairpins on reverse transcription loop-mediated isothermal amplification detection of viral RNA. Analyst 143, 1924-1933, doi:10.1039/c7an01897e (2018).

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. Environmental Science & Technology Letters 7, 511-516, doi:10.1021/acs.estlett.0c00357 (2020).

Millman, A., Melamed, S., Amitai, G. & Sorek, R. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. Nat Microbiol 5, 1608-1615, doi:10.1038/s41564-020-0777-y (2020).

Miorin, L. et al. SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. Proc Natl Acad Sci U S A 117, 28344-28354, doi:10.1073/pnas.2016650117 (2020).

Mistry, J. et al. Pfam: The protein families database in 2021. Nucleic Acids Res 49, D412-D419, doi:10.1093/nar/gkaa913 (2021).

Mo, C. Y. et al. Type III-A CRISPR immunity promotes mutagenesis of staphylococci. Nature 592, 611-615, doi:10.1038/s41586-021-03440-3 (2021).

Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733-740, doi:10.1099/mic.0.023960-0 (2009).

Mojica, F. J., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution 60, 174-182 (2005).

Monttinen, H. A., Ravantti, J. J. & Poranen, M. M. Common Structural Core of Three-Dozen Residues Reveals Intersuperfamily Relationships. Mol Biol Evol 33, 1697-1710, doi:10.1093/molbev/msw047 (2016).

Moon, S. B., Kim, D. Y., Ko, J. H. & Kim, Y. S. Recent advances in the CRISPR genome editing tool set. Exp Mol Med 51, 1-11, doi:10.1038/s12276-019-0339-7 (2019).

Morgan, M. et al. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics 25, 2607-2608, doi:10.1093/bioinformatics/btp450 (2009).

Mouw, K. W. & Rice, P. A. Shaping the Borrelia burgdorferi genome: crystal structure and binding properties of the DNA-bending protein Hbb. Molecular Microbiology 63, 1319-1330, doi:10.1111/j.1365-2958.2007.05586.x (2007).

Moya-Beltran, A. et al. Evolution of Type IV CRISPR-Cas Systems: Insights from CRISPR Loci in Integrative Conjugative Elements of Acidithiobacillia. CRISPR J 4, 656-672, doi:10.1089/crispr.2021.0051 (2021).

Muth, D. et al. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Sci Rep 8, 15177, doi:10.1038/s41598-018-33487-8 (2018).

Nasef, M. et al. Regulation of cyclic oligoadenylate synthesis by the Staphylococcus epidermidis Cas10-Csm complex. RNA 25, 948-962, doi:10.1261/rna.070417.119 (2019).

Nelson, C. A., Pekosz, A., Lee, C. A., Diamond, M. S. & Fremont, D. H. Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein. Structure 13, 75-85, doi:10.1016/j.str.2004.10.010 (2005).

Nemudryi, A. et al. SARS-CoV-2 genomic surveillance identifies naturally occurring truncation of ORF7a that limits immune suppression. Cell Rep 35, 109197, doi:10.1016/j.celrep.2021.109197 (2021).

Niewoehner, O. et al. Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. Nature 548, 543-548, doi:10.1038/nature23467 (2017).

Niewoehner, O. et al. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature 548, 543-548, doi:10.1038/nature23467 (2017).

Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156, 935-949 (2014).

Nivala, J., Shipman, S. L. & Church, G. M. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. Nat Microbiol 3, 310-318, doi:10.1038/s41564-017-0097-z (2018).

Notomi, T. et al. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res 28, E63, doi:10.1093/nar/28.12.e63 (2000).

Nuñez, James K., Bai, L., Harrington, Lucas B., Hinder, Tracey L. & Doudna, Jennifer A. CRISPR Immunological Memory Requires a Host Factor for Specificity. Molecular Cell 62, 824-833, doi:10.1016/j.molcel.2016.04.027 (2016).

O'Brien, R. E. et al. Structural basis for assembly of non-canonical small subunits into type I-C Cascade. Nat Commun 11, 5931, doi:10.1038/s41467-020-19785-8 (2020).

Ofir, G. et al. Antiviral activity of bacterial TIR domains via signaling molecules that trigger cell death. bioRxiv, 2021.2001.2006.425286, doi:10.1101/2021.01.06.425286 (2021).

Ooi, K. H. et al. An engineered CRISPR-Cas12a variant and DNA-RNA hybrid guides enable robust and rapid COVID-19 testing. Nat Commun 12, 1739, doi:10.1038/s41467-021-21996-6 (2021).

Organization, W. H. (2020).

Osada, N. et al. The genome landscape of the african green monkey kidney-derived vero cell line. DNA Res 21, 673-683, doi:10.1093/dnares/dsu029 (2014).

Osawa, T., Inanaga, H. & Numata, T. Crystal structure of the Cmr2-Cmr3 subcomplex in the CRISPR-Cas RNA silencing effector complex. J Mol Biol 425, 3811-3823, doi:10.1016/j.jmb.2013.03.042 (2013).

Osuna, B. A. et al. Listeria phages induce Cas9 degradation to protect lysogenic genomes. BioRxiv, 787200 (2019).

Ow, D. S. W., Lee, D. Y., Tung, H. H. & Lin-Chao, S. Plasmid Regulation and Systems-Level Effects on Escherichia coli Metabolism. Systems Biology and Biotechnology of Escherichia Coli, 273-294, doi:10.1007/978-1-4020-9394-4_14 (2009).

Ozcan, A. et al. Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. Nature 597, 720-725, doi:10.1038/s41586-021-03886-5 (2021).

Pagès H, A. P., Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.52.0 (2019).

Paltiel, A. D., Zheng, A. & Walensky, R. P. Assessment of SARS-CoV-2 Screening Strategies to Permit the Safe Reopening of College Campuses in the United States. JAMA Netw Open 3, e2016818, doi:10.1001/jamanetworkopen.2020.16818 (2020).

Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

Pati, A., Padhi, S., Suvankar, S. & Panda, A. K. Minor Allele of Interferon-Induced Transmembrane Protein 3 Polymorphism (rs12252) Is Covered Against Severe Acute Respiratory Syndrome Coronavirus 2 Infection and Mortality: A Worldwide Epidemiological Investigation. J Infect Dis 223, 175-178, doi:10.1093/infdis/jiaa630 (2021).

Pausch, P. et al. Structural Variation of Type I-F CRISPR RNA Guided DNA Surveillance. Mol Cell 67, 622-632 e624, doi:10.1016/j.molcel.2017.06.036 (2017).

Pawluk, A., Bondy-Denomy, J., Cheung, V. H., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. MBio 5, e00896, doi:10.1128/mBio.00896-14 (2014).

Pawluk, A., Davidson, A. R. & Maxwell, K. L. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol 16, 12 (2018).

Pawluk, A. et al. Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nat Microbiol 1, 16085, doi:10.1038/nmicrobiol.2016.85 (2016).

Peccia, J. et al. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat Biotechnol 38, 1164-1167, doi:10.1038/s41587-020-0684-z (2020).

Peng, R. et al. Structural insight into multistage inhibition of CRISPR-Cas12a by AcrVA4. Proceedings of the National Academy of Sciences 116, 18928-18936 (2019).

Peng, R. et al. Alternate binding modes of anti-CRISPR viral suppressors AcrF1/2 to Csy surveillance complex revealed by cryo-EM structures. Cell Res 27, 853-864, doi:10.1038/cr.2017.79 (2017).

Pérez-Martín, J., Rojo, F. & de Lorenzo, V. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. Microbiological Reviews 58, 268-290, doi:10.1128/mr.58.2.268-290.1994 (1994).

Petassi, M. T., Hsieh, S.-C. & Peters, J. E. Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons. Cell 183, 1757-1771.e1718, doi:10.1016/j.cell.2020.11.005 (2020).

Peters, J. E. Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond. Mol Microbiol 112, 1635-1644, doi:10.1111/mmi.14383 (2019).

Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proceedings of the National Academy of Sciences of the United States of America 114, E7358-E7366, doi:10.1073/pnas.1709035114 (2017).

Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30, 70-82, doi:10.1002/pro.3943 (2021).

Plagens, A., Richter, H., Charpentier, E. & Randau, L. DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. FEMS Microbiol Rev 39, 442-463 (2015).

Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653-663, doi:10.1099/mic.0.27437-0 (2005).

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 research 48, D535-D544, doi:10.1093/nar/gkz915 (2020).

Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490-e9490, doi:10.1371/journal.pone.0009490 (2010).

Pritchard, C. G., Weaver, D. T., Baril, E. F. & DePamphilis, M. L. DNA polymerase alpha cofactors C1C2 function as primer recognition proteins. J Biol Chem 258, 9810-9819 (1983).

Przybilski, R. et al. Csy4 is responsible for CRISPR RNA processing inPectobacterium atrosepticum. RNA Biology 8, 517-528, doi:10.4161/rna.8.3.15190 (2011).

Pul, Ü. et al. Identification and characterization ofE. coliCRISPR-caspromoters and their silencing by H-NS. Molecular Microbiology 75, 1495-1512, doi:10.1111/j.1365-2958.2010.07073.x (2010).

Putnam, C. D. & Tainer, J. A. Protein mimicry of DNA and pathway regulation. DNA Repair (Amst) 4, 1410-1420, doi:10.1016/j.dnarep.2005.08.007 (2005).

Qiu, D., Damron, F. H., Mima, T., Schweizer, H. P. & Yu, H. D. PBAD-based shuttle vectors for functional analysis of toxic and highly regulated genes in Pseudomonas and Burkholderia spp. and other bacteria. Appl Environ Microbiol 74, 7422-7426, doi:10.1128/AEM.01369-08 (2008).

Quick, J. in protocols.io (ZappyLab, Inc., 2020).

Quick, J. et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12, 1261-1276, doi:10.1038/nprot.2017.066 (2017).

Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228-232, doi:10.1038/nature16996 (2016).

Rambaut, A. & Drummond, A. (2012).

Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).

Randazzo, W. et al. SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. Water Res 181, 115942-115942, doi:10.1016/j.watres.2020.115942 (2020).

Rice, P. A., Yang, S.-w., Mizuuchi, K. & Nash, H. A. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. Cell 87, 1295-1306, doi:10.1016/s0092-8674(00)81824-3 (1996).

Richter, C. et al. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res 42, 8516-8526, doi:10.1093/nar/gku527 (2014).

Richter, C. et al. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic acids research 42, 8516-8526, doi:10.1093/nar/gku527 (2014).

Richter, C., Gristwood, T., Clulow, J. S. & Fineran, P. C. In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype IF CRISPR/Cas System. PLoS One 7, e49549 (2012).

Richter, C., Gristwood, T., Clulow, J. S. & Fineran, P. C. In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System. PLoS One 7, e49549-e49549, doi:10.1371/journal.pone.0049549 (2012).

Rimoldi, S. G. et al. Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. Sci Total Environ 744, 140911-140911, doi:10.1016/j.scitotenv.2020.140911 (2020).

Rolando, J. C., Jue, E., Barlow, J. T. & Ismagilov, R. F. Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification. Nucleic Acids Res 48, e42, doi:10.1093/nar/gkaa099 (2020).

Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic acids research 46, 1007-1020, doi:10.1093/nar/gkx1232 (2018).

Rollins, M. F. et al. Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. Mol Cell 74, 132-142 e135, doi:10.1016/j.molcel.2019.02.001 (2019).

Rollins, M. F. et al. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proceedings of the National Academy of Sciences of the United States of America 114, E5113-E5121, doi:10.1073/pnas.1616395114 (2017).

Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. & Wiedenheft, B. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res 43, 2216-2222 (2015).

Romero, G., von Zastrow, M. & Friedman, P. A. Role of PDZ proteins in regulating trafficking, signaling, and function of GPCRs: means, motif, and opportunity. Adv Pharmacol 62, 279-314, doi:10.1016/B978-0-12-385952-5.00003-8 (2011).

Rosenthal, S. H. et al. Identification of eight SARS-CoV-2 ORF7a deletion variants in 2,726 clinical specimens. bioRxiv, 2020.2012.2010.418855, doi:10.1101/2020.12.10.418855 (2020).

Rouillon, C., Athukoralage, J. S., Graham, S., Gruschow, S. & White, M. F. Control of cyclic oligoadenylate synthesis in a type III CRISPR system. Elife 7, doi:10.7554/eLife.36734 (2018).

Rouillon, C. et al. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell 52, 124-134, doi:10.1016/j.molcel.2013.08.020 (2013).

Rozkov, A. et al. Characterization of the metabolic burden on Escherichia coli DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. Biotechnol Bioeng 88, 909-915, doi:10.1002/bit.20327 (2004).

Rybarski, J. R. H., Kuang; Hill, Alexis M.; Wilke, Claus O. ; Finkelstein, Ilya J. Metagenomic Discovery of CRISPR-Associated Transposons. bioRxiv, doi:10.1101/2021.08.16.456562 (2021).

Salah Ud-Din, A. I., Tikhomirova, A. & Roujeinikova, A. Structure and Functional Diversity of GCN5-Related N-Acetyltransferases (GNAT). Int J Mol Sci 17, doi:10.3390/ijms17071018 (2016).

Samai, P. et al. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell 161, 1164-1174, doi:10.1016/j.cell.2015.04.027 (2015).

Santiago-Frangos, A. et al. Intrinsic signal amplification by type III CRISPR-Cas systems provides a sequence-specific SARS-CoV-2 diagnostic. Cell Rep Med 2, 100319, doi:10.1016/j.xcrm.2021.100319 (2021).

Santiago-Frangos, A., Wiegand, T. & Wiedenheft, B. Cas9 slide-and-seek for phage defense and genome engineering. EMBO J 38, doi:10.15252/embj.2019101474 (2019).

Schleif, R. DNA LOOPING. Annual Review of Biochemistry 61, 199-223, doi:10.1146/annurev.bi.61.070192.001215 (1992).

Schliep, K. P. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. Nature 562, 380-385, doi:10.1038/s41586-018-0569-1 (2018).

Schoggins, J. W. Interferon-Stimulated Genes: What Do They All Do? Annu Rev Virol 6, 567-584, doi:10.1146/annurev-virology-092818-015756 (2019).

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 108, 10098-10103, doi:10.1073/pnas.1104144108 (2011).

Shi, G. et al. Opposing activities of IFITM proteins in SARS-CoV-2 infection. EMBO J 40, e106501, doi:10.15252/embj.2020106501 (2021).

Shibata, M. et al. Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. Nat Commun 8, 1430, doi:10.1038/s41467-017-01466-8 (2017).

Shin, J. et al. Disabling Cas9 by an anti-CRISPR DNA mimic. Sci Adv 3, e1701620, doi:10.1126/sciadv.1701620 (2017).

Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. Science 353, aaf1175-aaf1175, doi:10.1126/science.aaf1175 (2016).

Shmakov, S. et al. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res 42, 5907-5916, doi:10.1093/nar/gku226 (2014).

Silverman, J. D., Hupert, N. & Washburne, A. D. Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. Sci Transl Med 12, eabc1126, doi:10.1126/scitranslmed.abc1126 (2020).

Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. Nat Commun 7, 12778, doi:10.1038/ncomms12778 (2016).

Sofos, N. et al. Structures of the Cmr-beta Complex Reveal the Regulation of the Immunity Mechanism of Type III-B CRISPR-Cas. Mol Cell 79, 741-757 e747, doi:10.1016/j.molcel.2020.07.008 (2020).

Songailiene, I. et al. HEPN-MNT Toxin-Antitoxin System: The HEPN Ribonuclease Is Neutralized by OligoAMPylation. Mol Cell 80, 955-970 e957, doi:10.1016/j.molcel.2020.11.034 (2020).

Sridhara, S., Goswami, H. N., Whyms, C., Dennis, J. H. & Li, H. Virus detection via programmable Type III-A CRISPR-Cas systems. Nat Commun 12, 5653, doi:10.1038/s41467-021-25977-7 (2021).

Sridhara, S. et al. Structure and Function of an in vivo Assembled Type III-A CRISPR-Cas Complex Reveal Critical Roles of Dynamics in Activity Control. bioRxiv, 2021.2001.2027.428455, doi:10.1101/2021.01.27.428455 (2021).

Staals, R. H. et al. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun 7, 12853, doi:10.1038/ncomms12853 (2016).

Staals, R. H. et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol Cell 56, 518-530, doi:10.1016/j.molcel.2014.10.005 (2014).

Stanley, S. Y. et al. Anti-CRISPR-Associated Proteins Are Crucial Repressors of Anti-CRISPR Transcription. Cell 178, 1452-1464. e1413 (2019).

Stanley, S. Y. & Maxwell, K. L. Phage-Encoded Anti-CRISPR Defenses. Annu Rev Genet 52, 445-464, doi:10.1146/annurev-genet-120417-031321 (2018).

Steens, J. A. et al. SCOPE enables type III CRISPR-Cas diagnostics using flexible targeting and stringent CARF ribonuclease activation. Nat Commun 12, 5033, doi:10.1038/s41467-021-25337-5 (2021).

Stefanelli, P. et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. Euro Surveill 25, 2000305, doi:10.2807/1560-7917.ES.2020.25.13.2000305 (2020).

Stella, S. et al. Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. Cell 175, 1856-1871 e1821, doi:10.1016/j.cell.2018.10.045 (2018).

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507, 62 (2014).

Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR-Cas systems. Mol Cell 61, 797-808 (2016).

Stoltz, M. & Klingstrom, J. Alpha/beta interferon (IFN-alpha/beta)-independent induction of IFN-lambda1 (interleukin-29) in response to Hantaan virus infection. J Virol 84, 9140-9148, doi:10.1128/JVI.00717-10 (2010).

Stone, N. P. et al. A hyperthermophilic phage decoration protein suggests common evolutionary origin with herpesvirus triplex proteins and an anti-CRISPR protein. Structure 26, 936-947. e933 (2018).

Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. Science 365, 48-53, doi:10.1126/science.aax9181 (2019).

Su, Y. C. F. et al. Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. mBio 11, doi:10.1128/mBio.01610-20 (2020).

Sun, W. et al. Structures of Neisseria meningitidis Cas9 Complexes in Catalytically Poised and Anti-CRISPR-Inhibited States. Mol Cell, doi:10.1016/j.molcel.2019.09.025 (2019).

Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. Current Opinion in Structural Biology 14, 28-35, doi:10.1016/j.sbi.2003.12.003 (2004).

Tamulaitis, G. et al. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell 56, 506-517, doi:10.1016/j.molcel.2014.09.027 (2014).

Tang, A. et al. Detection of Novel Coronavirus by RT-PCR in Stool Specimen from Asymptomatic Child, China. Emerg Infect Dis 26, 1337-1339, doi:10.3201/eid2606.200301 (2020).

Tang, J., Akerboom, J., Vaziri, A., Looger, L. L. & Shank, C. V. Near-isotropic 3D optical nanoscopy with photon-limited chromophores. Proc Natl Acad Sci U S A 107, 10068-10073, doi:10.1073/pnas.1004899107 (2010).

Tay, M. Z., Poh, C. M., Renia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. Nat Rev Immunol 20, 363-374, doi:10.1038/s41577-020-0311-8 (2020).

Thavalingam, A. et al. Inhibition of CRISPR-Cas9 ribonucleoprotein complex assembly by anti-CRISPR AcrIIC2. Nat Commun 10, 2806, doi:10.1038/s41467-019-10577-3 (2019).

Thi Nhu Thao, T. et al. Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. Nature 582, 561-565, doi:10.1038/s41586-020-2294-9 (2020).

Tomita, N., Mori, Y., Kanda, H. & Notomi, T. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. Nat Protoc 3, 877-882, doi:10.1038/nprot.2008.57 (2008).

Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv, doi:10.1101/2020.09.04.283077 (2020).

Vale, P. F. & Little, T. J. CRISPR-mediated phage resistance and the ghost of coevolution past. Proc Biol Sci 277, 2097-2103, doi:10.1098/rspb.2010.0055 (2010).

van Dijk, M. & Bonvin, A. M. J. J. 3D-DART: a DNA structure modelling server. Nucleic acids research 37, W235-W239, doi:10.1093/nar/gkp287 (2009).

van Dorp, L. et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83, 104351-104351, doi:10.1016/j.meegid.2020.104351 (2020).

van Erp, P. B. G. et al. Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. ACS Chem Biol 13, 481-490, doi:10.1021/acschembio.7b00649 (2018).

Vink, J. N. A. et al. Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. Mol Cell, doi:10.1016/j.molcel.2019.10.021 (2019).

Vlot, M. et al. Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. Nucleic acids research 46, 873-885 (2017).

Volz, E. et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv, 2020.2012.2030.20249034, doi:10.1101/2020.12.30.20249034 (2021).

Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res 43, 10848-10860, doi:10.1093/nar/gkv1261 (2015).

Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic acids research 43, 10848-10860, doi:10.1093/nar/gkv1261 (2015).

Wang, B., Xu, W. & Yang, H. Structural basis of a Tn7-like transposase recruitment and DNA loading to CRISPR-Cas surveillance complex. Cell Res, doi:10.1038/s41422-020-0274-0 (2020).

Wang, C. et al. The Coronavirus PEDV Evades Type III Interferon Response Through the miR-30c-5p/SOCS1 Axis. Front Microbiol 11, 1180, doi:10.3389/fmicb.2020.01180 (2020).

Wang, J. et al. A CRISPR evolutionary arms race: structural insights into viral anti-CRISPR/Cas responses. Cell Res 26, 1165 (2016).

Wang, J. C. Helical repeat of DNA in solution. Proceedings of the National Academy of Sciences of the United States of America 76, 200-203, doi:10.1073/pnas.76.1.200 (1979).

Wang, W. et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA 323, 1843-1844, doi:10.1001/jama.2020.3786 (2020).

Wang, X. et al. Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. Nat Struct Mol Biol 23, 868-870, doi:10.1038/nsmb.3269 (2016).

Watters, K. E., Fellmann, C., Bai, H. B., Ren, S. M. & Doudna, J. A. Systematic discovery of natural CRISPR-Cas12a inhibitors. Science 362, 236-239 (2018).

Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic acids research 43, 1749-1758, doi:10.1093/nar/gku1407 (2015).

Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev 29, 356-361, doi:10.1101/gad.257550.114 (2015).

Westra, Edze R. et al. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. Current Biology 25, 1043-1049, doi:10.1016/j.cub.2015.01.065 (2015).

Whiteley, A. T. et al. Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. Nature 567, 194-199, doi:10.1038/s41586-019-0953-5 (2019).

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. R package version 3.2.1 (2019).

Wickham, H. ggpubr: ‘ggplot2' Based Publication Ready Plots. R package version 0.2.3 (2019).

Wickham, H. et al. Welcome to the Tidyverse. Journal of Open Source Software 4, 1686, doi:10.21105/joss.01686 (2019).

Wiedenheft, B. et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A 108, 10092-10097, doi:10.1073/pnas.1102716108 (2011).

Wiegand, T., Karambelkar, S., Bondy-Denomy, J. & Wiedenheft, B. Structures and Strategies of Anti-CRISPR-Mediated Immune Suppression. Annu Rev Microbiol 74, 21-37, doi:10.1146/annurev-micro-020518-120107 (2020).

Wiegand, T. et al. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. The CRISPR journal 3, 378-387, doi:10.1089/crispr.2020.0069 (2020).

Wiegand, T. & Wiedenheft, B. CRISPR Surveillance Turns Transposon Taxi. CRISPR J 3, 10-12, doi:10.1089/crispr.2020.29081.twi (2020).

Wilkinson, M. et al. Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. Mol Cell 75, 90-101 e105, doi:10.1016/j.molcel.2019.04.020 (2019).

Wolfel, R. et al. Virological assessment of hospitalized patients with COVID-2019. Nature 581, 465-469, doi:10.1038/s41586-020-2196-x (2020).

Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. Nature Structural & Molecular Biology 23, 876-883, doi:10.1038/nsmb.3289 (2016).

Wright, A. V. et al. Structures of the CRISPR genome integration complex. Science 357, 1113-1118, doi:10.1126/science.aao0679 (2017).

Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics 16, 322, doi:10.1186/s12859-015-0749-z (2015).

Wu, F. et al. SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. mSystems 5, doi:10.1128/mSystems.00614-20 (2020).

Wu, F. et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

Wu, Y. et al. Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol Hepatol 5, 434-435, doi:10.1016/S2468-1253(20)30083-2 (2020).

Wurtzer, S. et al. Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, Greater Paris, France, 5 March to 23 April 2020. Euro Surveill 25, doi:10.2807/1560-7917.ES.2020.25.50.2000776 (2020).

Xia, H. et al. Evasion of Type I Interferon by SARS-CoV-2. Cell Rep 33, 108234, doi:10.1016/j.celrep.2020.108234 (2020).

Xiao, F. et al. Infectious SARS-CoV-2 in Feces of Patient with Severe COVID-19. Emerg Infect Dis 26, 1920-1922, doi:10.3201/eid2608.200681 (2020).

Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science 361, doi:10.1126/science.aat0839 (2018).

Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature 550, 137-141, doi:10.1038/nature24020 (2017).

Xie, X. et al. An Infectious cDNA Clone of SARS-CoV-2. Cell Host Microbe 27, 841-848 e843, doi:10.1016/j.chom.2020.04.004 (2020).

Xu, Y. et al. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat Med 26, 502-505, doi:10.1038/s41591-020-0817-4 (2020).

Yan, W. X. et al. Functionally diverse type V CRISPR-Cas systems. Science 363, 88-91, doi:10.1126/science.aav7271 (2019).

Yang, H. & Patel, D. J. Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. Mol Cell 67, 117-127 e115, doi:10.1016/j.molcel.2017.05.024 (2017).

Yang, M. et al. The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. Cell Rep 22, 372-382, doi:10.1016/j.celrep.2017.12.048 (2018).

Yang, W. An overview of Y-Family DNA polymerases and a case study of human DNA polymerase eta. Biochemistry 53, 2793-2803, doi:10.1021/bi500019s (2014).

Yao, J. et al. Novel polyadenylylation-dependent neutralization mechanism of the HEPN/MNT toxin/antitoxin system. Nucleic Acids Res 48, 11054-11067, doi:10.1093/nar/gkaa855 (2020).

Yao, S., Helinski, D. R. & Toukdarian, A. Localization of the naturally occurring plasmid ColE1 at the cell pole. J Bacteriol 189, 1946-1953, doi:10.1128/JB.01451-06 (2007).

Ye, Y., Ellenberg, R. M., Graham, K. E. & Wigginton, K. R. Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater. Environ Sci Technol 50, 5077-5085, doi:10.1021/acs.est.6b00876 (2016).

Yoganand, K. N. R., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic acids research 45, 367-381, doi:10.1093/nar/gkw1151 (2017).

Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic acids research 40, 5569-5576, doi:10.1093/nar/gks216 (2012).

You, L. et al. Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. Cell 176, 239-253 e216, doi:10.1016/j.cell.2018.10.052 (2019).

Young, B. E. et al. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet 396, 603-611, doi:10.1016/S0140-6736(20)31757-8 (2020).

Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. Current Protocols in Bioinformatics 69, doi:10.1002/cpbi.96 (2020).

Yu, G., Lam, T. T., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol Biol Evol 35, 3041-3043, doi:10.1093/molbev/msy194 (2018).

Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol Biol Evol 35, 3041-3043, doi:10.1093/molbev/msy194 (2018).

Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8, 28-36, doi:10.1111/2041-210x.12628 (2016).

Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284-287, doi:10.1089/omi.2011.0118 (2012).

Zabrady, K., Zabrady, M., Kolesar, P., Li, A. W. H. & Doherty, A. J. CRISPR-Associated Primase-Polymerases are implicated in prokaryotic CRISPR-Cas adaptation. Nat Commun 12, 3690, doi:10.1038/s41467-021-23535-9 (2021).

Zang, R. et al. Cholesterol 25-hydroxylase suppresses SARS-CoV-2 replication by blocking membrane fusion. Proc Natl Acad Sci U S A 117, 32105-32113, doi:10.1073/pnas.2012197117 (2020).

Zehender, G. et al. Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. J Med Virol 92, 1637-1640, doi:10.1002/jmv.25794 (2020).

Zhang, B. et al. Clinical characteristics of 82 cases of death from COVID-19. PLoS One 15, e0235458, doi:10.1371/journal.pone.0235458 (2020).

Zhang, H. et al. Structural Basis for the Inhibition of CRISPR-Cas12a by Anti-CRISPR Proteins. Cell Host Microbe 25, 815-826 e814, doi:10.1016/j.chom.2019.05.004 (2019).

Zhang, J., Graham, S., Tello, A., Liu, H. & White, M. F. Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. Nucleic Acids Res 44, 1789-1799, doi:10.1093/nar/gkw020 (2016).

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614-620, doi:10.1093/bioinformatics/btt593 (2014).

Zhang, S. et al. Mycobacterium tuberculosis CRISPR/Cas system Csm1 holds clues to the evolutionary relationship between DNA polymerase and cyclase activity. Int J Biol Macromol 170, 140-149, doi:10.1016/j.ijbiomac.2020.12.014 (2021).

Zhang, Y. et al. Interferon-Induced Transmembrane Protein 3 Genetic Variant rs12252-C Associated With Disease Severity in Coronavirus Disease 2019. J Infect Dis 222, 34-37, doi:10.1093/infdis/jiaa224 (2020).

Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 395, 1054-1062, doi:10.1016/S0140-6736(20)30566-3 (2020).

Zhou, Y. et al. Structure of a type IV CRISPR-Cas ribonucleoprotein complex. iScience 24, 102201, doi:10.1016/j.isci.2021.102201 (2021).

Zhou, Z. et al. Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14(+) monocytes. iScience 24, 102187, doi:10.1016/j.isci.2021.102187 (2021).

Zhu, Y. et al. Diverse Mechanisms of CRISPR-Cas9 Inhibition by Type IIC Anti-CRISPR Proteins. Mol Cell 74, 296-309 e297, doi:10.1016/j.molcel.2019.01.038 (2019).

Zimmermann, L. et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430, 2237-2243, doi:10.1016/j.jmb.2017.12.007 (2018).

Zou, Y., Mason, M. G. & Botella, J. R. Evaluation and improvement of isothermal amplification methods for point-of-need plant disease diagnostics. PLoS One 15, e0235216, doi:10.1371/journal.pone.0235216 (2020).