

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Genomic hallmarks of self and non-self in antiviral immunity

Permalink

<https://escholarship.org/uc/item/1sg9s208>

ISBN

9798189270215

Author

Tuck, Owen Templeton

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Genomic hallmarks of *self* and *non-self* in antiviral immunity

By

Owen Templeton Tuck

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jennifer A. Doudna, Chair

Professor Michael A. Marletta

Professor James K. Nuñez

Spring 2026

Genomic hallmarks of *self* and *non-self* in antiviral immunity

Copyright © 2026

By

Owen Templeton Tuck

Abstract

Genomic hallmarks of *self* and non-self in antiviral immunity

By

Owen Templeton Tuck

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Jennifer A. Doudna, Chair

Professor Michael A. Marletta

Professor James K. Nuñez

Viruses are fundamental drivers of cellular evolution. Myriad immune systems evolved to combat viral infections, but these systems face a fundamental challenge: distinguishing *self* from *non-self*. Immune systems must mount a potent response to infection, but immunity must be specific and target only pathogens or infected cells. *Nonself* immune responses run the dire risk of autoimmunity. How do organisms evolve specific immune systems? This work explores two widespread immune systems in prokaryotes, CRISPR-Cas and Hachiman, to understand how molecular solutions evolved to address the core constraint of autoimmunity. These studies reveal that DNA topology manipulation and genome integrity sensing represent connected mechanisms for specifically targeting *non-self*.

How do CRISPR-Cas adaptive systems avoid autoimmunity? CRISPR immune systems capture DNA fragments from mobile genetic elements and integrate them into the host genome as templates for RNA-guided immunity. The Cas1-Cas2 integrase is responsible for driving adaptation and must therefore distinguish *self* from *non-self*, but the mechanism was unclear. My biochemical and structural analyses reveal that a Cas1–Cas2/exonuclease fusion (trimmer-integrase) performs intricate DNA topological manipulations to protect and remove a precursor motif (PAM) that serves as an immunological tag distinguishing self from *non-self*.

Most natural immune systems are innate rather than adaptive, and many recently discovered systems encode helicases. While helicases sense pathogen-associated nucleic acids in eukaryotic immunity, their prokaryotic roles remain understudied. I characterized the widespread Hachiman system using *in vivo* assays, biochemistry, cryo-EM, and phylogenetics, demonstrating it functions as a nuclease-helicase complex (HamAB). HamB acts as a sensor helicase that detects DNA damage, triggering allosteric activation of the HamA nuclease to degrade cellular DNA. DNA-damaging agents activate Hachiman without phage presence, indicating this system monitors genome integrity as a proxy for infection. I propose a thresholding model explaining how Hachiman-like systems avoid autoimmunity and achieve broad-spectrum defense.

In addition to core immunity mechanisms, I discovered regulatory layers controlling immune responses. Gene context analysis revealed coevolving regulatory gene pairs acquired by multiple immune systems that encode a specific protease that activates a partner pro-nuclease to trigger genome destruction. This analysis was extended to identify caspase-nuclease (*canu*) family systems. *Canu* confers antiviral defense through a pathway reminiscent of eukaryotic caspase activation. This study demonstrates how modularity and effector exchange drive immune evolution and sets the stage to uncover key events in caspase evolution.

These findings establish genome integrity sensing as a paradigm of prokaryotic immunity with broad evolutionary implications. The Hachiman helicase shares ancestry with critical eukaryotic helicases including Ski2 and Brr2, suggesting origins of eukaryotic regulatory strategies in immunity. Modular evolution revealed through nuclease-protease pairs and *canu* systems demonstrates how immune components evolve as interchangeable units, accelerating adaptation and presenting a new avenue for discovery. Our ability to trigger immune activation without viral presence challenges traditional models of immune specificity and reveals these systems may monitor cellular homeostasis rather than directly detecting pathogens. This work establishes molecular signatures of *non-self* in the genome across diverse immune systems and informs our understanding how antiviral defenses function and evolve.

To my parents, Suzanne and Andy Tuck.

Table of Contents

Table of Contents.....	4
List of Figures.....	24
Acknowledgments.....	25
Chapter 1 Introduction.....	1
1.1 <i>Self</i> and <i>non-self</i>	2
1.3 Ancestral immunity.....	4
1.4 Bacteriophages.....	5
1.5 Pan-immunity.....	5
Chapter 2 Genome expansion by a CRISPR trimmer-integrase.....	10
2.1 Abstract.....	11
2.2 Introduction.....	11
2.3 Results.....	12
2.3.1 Cas1:Cas2-DEDDh substrate generation.....	12
2.3.2 Cas1 binds the PAM and blocks DEDDh trimming during prespacer processing.....	14
2.3.3 DEDDh trims the PAM after DNA half-integration.....	16
2.3.4 Reconstitution of full CRISPR array integration.....	19
2.3.5 Selective processing and protection of prespacer sources and models of biogenesis.....	21
2.4 Discussion.....	23
2.5 Methods.....	25
2.5.1 Plasmid construction and DNA substrate preparation.....	25
2.5.2 Cloning, expression, and purification.....	26
2.5.3 Processing assays.....	26
2.5.4 Cryo-EM data acquisition.....	26
2.5.5 Cryo-EM data processing.....	27
2.5.6 Model building and refinement.....	28
2.5.7 Ligation assays with pCRISPR integration target plasmid.....	29
2.5.8 Full-site integration assays.....	29
2.5.9 CRISPR locus bioinformatic analysis.....	30
2.5.10 Statistics and reproducibility.....	30
2.5.11 Data availability.....	30
2.6 Supplementary Information.....	31

Chapter 3 Genome integrity sensing by the broad-spectrum Hachiman antiphage defense complex.....	46
3.1 Abstract.....	47
3.2 Introduction.....	47
3.3 Results.....	48
3.3.1 Hachiman is broad-spectrum helicase defense system.....	48
3.3.2 Structural basis of HamAB complexation.....	50
3.3.3 HamA DUF1837 encodes a nuclease.....	52
3.3.4 HamB is a DNA helicase	52
3.3.5 HamAB degrades plasmids in vitro.....	54
3.3.6 Structural basis of HamB DNA binding.....	55
3.3.7 Helicase ratcheting may activate HamAB	56
3.3.8 Hachiman degrades phage and host DNA simultaneously during infection.....	57
3.3.9 DNA damage activates Hachiman	58
3.3.10 Hachiman associates with intact dsDNA.....	60
3.4 Discussion.....	62
3.5 Methods.....	65
3.5.1 Bacterial strains and bacteriophages.....	65
3.5.2 Helicase and Hachiman phylogenetic analysis.....	66
3.5.3 Identification and visualization of Hachiman-containing loci.....	67
3.5.4 Plasmid and strain construction.....	67
3.5.5 Plaque assays.....	67
3.5.6 Bacteriophage liquid growth assays.....	68
3.5.7 Bacterial liquid growth assays.....	68
3.5.8 DNA Substrate Preparation.....	69
3.5.9 Protein expression and purification.....	69
3.5.10 Cryo-EM sample preparation and data acquisition.....	70
3.5.12 Cryo-EM data processing.....	70
3.5.13 Model building.....	72
3.5.14 NTPase assays.....	72
3.5.15 HamAB activity assays.....	73
3.5.16 Gel-shift helicase unwinding assays and quantification.....	73
3.5.17 Live single-cell time-lapse and static fluorescence microscopy.....	74
3.5.18 Quantification and statistical analysis.....	74
3.5.19 Data and code availability.....	75

3.6 Supplementary Information.....	76
Chapter 4 Recurrent acquisition of nuclease-protease pairs in antiviral immunity.....	92
4.1 Abstract.....	93
4.2 Introduction.....	93
4.3 Results.....	94
4.3.1 Repeated acquisition of paired genes in multiple antiphage systems..	94
4.3.2 Hachiman-regulated proteolytic cleavage unmasks the nuclease active site.....	95
4.3.3 DNA sensing and oligomerization trigger HamMAB defense.....	98
4.3.4 Gated nuclease activation is conserved in AVAST.....	99
4.3.5 Identification of a caspase-nuclease defense system.....	101
4.4 Discussion.....	104
4.5 Methods.....	105
4.5.1 Nuclease phylogenetic and coevolutionary analyses.....	105
4.5.2 Hachiman HamA phylogenetic analysis.....	107
4.5.3 Bacterial strains and phages.....	107
4.5.4 Plasmid construction.....	108
4.5.5 Plaque assays.....	108
4.5.6 Liquid infection assays.....	109
4.5.7 Protein expression and purification.....	109
4.5.8 Protease-HamAB proteolytic cleavage assays.....	110
4.5.9 In-gel mass spectrometry.....	110
4.5.10 Plasmid and phage gDNA clearance assays.....	111
4.5.11 Protease-HamAB nuclease short nucleic acid activity assay.....	112
4.5.12 Malachite green ATPase assays.....	112
4.5.13 Nuclease and caspase structural analyses.....	112
4.5.14 Liquid toxicity assays.....	113
4.5.15 Western blot analyses.....	113
4.5.16 Isolation, scaling validation and sequencing of mutant phages that overcome defense.....	114
4.6 Supplementary Information.....	115
References.....	136

List of Figures

Figure 2-1. Cas1–Cas2/DEDDh processes prespacers to the correct size for integration and protects a TT PAM.....	13
Figure 2-2. Molecular detail of Cas1–Cas2/DEDDh during prespacer processing.....	15
Figure 2-3. Biochemical and structural analysis of Cas1–Cas2/DEDDh PAM processing.....	17
Figure 2-4. In vitro reconstitution of Cas1:Cas2-DEDDh-mediated full-site integration.....	20
Figure 2-5. Cas1:Cas2-DEDDh selectively performs PAM-dependent trimming on certain ssDNA substrates but not dsDNA.....	22
Supplementary Figure 2-1. Cas1 – Cas2/DEDDh exhibits strict requirement for pre-duplexed 23-bp DNA duplex prespacer for functional ruler-guided trimming.....	32
Supplementary Figure 2-2. Kinetics of exonuclease trimming and ligation by Cas1:Cas2-DEDDh in vitro.....	33
Supplementary Figure 2-3. Flow-chart of the cryo-EM single particle reconstruction of the PAM-deficient prespacer bound Cas1-Cas2/DEDDh.....	35
Supplementary Figure 2-4. Flow-chart of the cryo-EM single particle reconstruction of the PAM-containing prespacer bound Cas1:Cas2 and resolution of the DEDDh density.....	37
Supplementary Figure 2-5. Flow-chart of the cryo-EM single particle reconstruction of the PAM-containing half-site substrate bound to Cas1:Cas2-DEDDh.....	39
Supplementary Figure 2-6. Incomplete resistance of phosphorothioate PAM nucleotides to DEDDh.....	40
Supplementary Figure 2-7. Molecular detail of Cas1:Cas2-DEDDh in the linear CRISPR repeat DNA conformation.....	41
Supplementary Figure 2-8. Integration site preference, orientation bias, and duplex requirements of Cas1-Cas2:DEDDh.....	42
Supplementary Figure 2-9. Comparison of Cas1:Cas2-DEDDh bound to the half-site intermediate with the complete <i>E. coli</i> IHF-containing integration complex.....	44
Supplementary Figure 2-10. Comparison of Cas1:Cas2-DEDDh to <i>E. coli</i> Cas1:Cas2 structures.....	45
Figure 3-1. Hachiman is a two-component defense system that protects against diverse bacteriophages.....	49
Figure 3-2. Structural basis of Hachiman complexation and identification of the HamA active site.....	51
Figure 3-3. HamAB is a DNA nuclease/helicase that degrades plasmids <i>in vitro</i>	53
Figure 3-4. Structural basis of HamB-DNA binding and helicase ratcheting.....	56

Figure 3-5. Hachiman defends against bacteriophage by nonspecific DNA clearance.....	58
Figure 3-6. DNA damage activates Hachiman.....	60
Figure 3-7. Hachiman scans intact dsDNA.....	62
Supplementary Figure 3-1. Helicases and Hachiman phylogenetic analysis.....	77
Supplementary Figure 3-2. Function and phylogeny of tested HamAB proteins.....	78
Supplementary Figure 3-3. Cryo-EM structure of the HamAB apo complex.....	80
Supplementary Figure 3-4. Analysis of HamAB biochemical activities in vitro.....	83
Supplementary Figure 3-5. Cryo-EM structure of two HamB-DNA complex conformations.....	84
Supplementary Figure 3-6. Drug-induced DNA damage activates Hachiman.....	87
Supplementary Figure 3-7. Cryo-EM structure of a HamA(E138A, K140A)B-plasmid DNA complex.....	89
Figure 4-1. Multiple defense systems acquired a predicted nuclease and fused protease.....	94
Figure 4-2. HamMAB encodes a specific protease which activates a partner nuclease.....	97
Figure 4-3. Nuclease-protease pairs function in AVAST.....	100
Figure 4-4. <i>Canu</i> is an antiphage defense system with a caspase-activated nuclease.....	102
Supplementary Figure 4-1. Structural comparison of nuclease-protease pairs in known defense systems.....	115
Supplementary Figure 4-2. Modular evolution of defense-associated MBL hydrolase-like nucleases.....	117
Supplementary Figure 4-3. Comparison between species and nuclease trees and reconciliation analysis.....	118
Supplementary Figure 4-4. Comparison of nuclease and trypsin-like protease phylogenies.....	119
Supplementary Figure 4-5. Nuclease-protease Hachiman systems are highly diverse.....	120
Supplementary Figure 4-6. Hachiman HamA acquired multiple additional domains.....	121
Supplementary Figure 4-7. HamA domain phylogeny and loss of nuclease activity.....	123
Supplementary Figure 4-8. <i>Pseudomonas fluorescens</i> HamMAB is antiphage.....	126
Supplementary Figure 4-9. Plaque assays of <i>P. fluorescens</i> HamMAB mutants.....	126
Supplementary Figure 4-10. Complete liquid infection data for <i>Pfl</i> /HamMAB.....	127
Supplementary Figure 4-11. Purification of HamM and HamAB.....	128

Supplementary Figure 4-12. <i>In vitro</i> incubation of <i>Pfl</i> HamAB with Single-Stranded DNA Binding Protein (SSB), RecBCD, and Topoisomerase I.....	128
Supplementary Figure 4-13. <i>In vitro</i> cleavage of various nucleic acids by <i>Pfl</i> HamMAB.....	129
Supplementary Figure 4-14. Pre-cut HamM is constitutively toxic to host cells.....	129
Supplementary Figure 4-15. Drug-induced <i>Pfl</i> HamMAB toxicity.....	130
Supplementary Figure 4-16. <i>In vitro</i> ATPase assay of <i>Pfl</i> HamAB.....	131
Supplementary Figure 4-17. <i>Epi</i> Avs liquid infection and terminase-mediated activation.....	132
Supplementary Figure 4-18. Plaque assays of <i>E. piriflorinigra</i> ns AVAST and Avs1a mutants.....	133
Supplementary Figure 4-19. Structural comparisons of trypsin-associated nucleases.....	134
Supplementary Figure 4-20. Plaque assays of <i>K. aerogenes</i> CanABC and mutants.....	134
Supplementary Figure 4-21. Phage escape of <i>Kae</i> CanABC involves mutation of the <i>dda</i> helicase.....	135

Acknowledgments

How do I even begin to thank the people who made this possible? To express my gratitude, I should first explain how I've come to view research as a whole. The way I see it, research seeks a very special kind of knowledge. You can't find this knowledge in books or at lectures, nor can it be had by passive observation. Rather, the objective of research is an active knowledge borne out of *experience*. To learn something new, we find a way to experience the problem, to fuss with it and squish it and add new dimension to it, probably breaking it along the way. I find this process extremely fulfilling and incredibly fun. It speaks to a deep desire in me to understand by doing. The thing about research that I hadn't fully appreciated before grad school is that, at its core, great research is deeply *social*. I relied so much on my mentors and colleagues to learn how to research effectively, and they nurtured my love for it and enhanced it immeasurably by their own contributions. The community of people around me is my source of inspiration - with them, I feel sustained even when science feels like a hopeless struggle. During my Ph.D., I got to live the life of research with them. I think I have found my calling.

From Jennifer Doudna, my advisor, I learned what it means to research with purpose and vision. Jennifer always insisted that I ask the simplest questions because they often hide the most difficult realities and biases: why are you doing this? Is this the right experiment? Do you really know what's going on here? Have you controlled for this or that possibility? Her incisive questions always come from a place of consideration, care for her mentees, and genuine curiosity, and they shaped the way I practice and share my research. Along with her steadfast mentorship and her brilliance, I want to thank Jennifer for her trust. After I finished my first project in the lab, Jennifer didn't ask me to continue studying adaptation or to generate new data. Instead she simply asked: "what do you want to do next?" Her trust meant and means a lot to me. I always felt it behind me as I went to ask new questions in new systems and experience exciting things that nourish me intellectually. In the future, I'm resolved to reciprocate the trust I received from Jennifer to my mentees as I continue on this path and hopefully build a lab of my own. Thank you Jennifer for everything.

Jennifer cultivates an amazing assortment of people with whom I had the privilege of working these past five years. When I joined the lab, Joy Wang was my experienced mentor and guide. Joy demystified the lab and was patient as I adjusted to thinking about things from a biochemist's perspective. I admire her clarity of thought and I'm deeply grateful to her for taking me on as a rotation student and showing me the wonderful and underappreciated world of CRISPR-Cas adaptation. During that initial period in the Doudna Lab, I was lucky to get the chance to learn cryogenic electron microscopy from Petr Skopintsev and Kasia Soczek. This opened up a new world for me - one where I could actually see the biomolecules I was playing around with in the

lab. Kasia is a wonderful friend to me and was a constant source of support. Julia Zhou, my former mentee and now aspiring doctor, was with me from the start. Looking back, I realize how gracious she was to put up with my wild ideas and still nascent understanding of things (we tried to do a simple PCR for two weeks). All that is a testament to her endless kindness and her amazing work ethic. I know Julia will be the finest doctor.

During a sort of sophomore slump, I started brainstorming with Ben Adler, our local phage expert and overall great human being. Soon, we started to work on a weird system called 'Hachiman'. I look back with fondness on the times when it seemed that every day we learned a new, essential bit of information and connected parts of the puzzle together. One night late in the lab, we came up with the key experiment and wrote it down frantically on a napkin that I still have in my desk drawer. Those days encapsulate the sheer fun of working with other people on a tricky problem and experiencing the great joy of discovery. I couldn't have asked for a better partner in Hachiman than Ben, and I'm proud of what we accomplished. Jason Hu was also an essential part of the Hachiman story, both part 1 and part 2. He joined as a determined rotation student and smashed all expectations of what one could accomplish in a mere 10 weeks. His contributions were essential, and he was a productive co-lead on the follow-up story. More than once, he corrected a key error I had made and developed the model in a new and insightful way. I'm grateful to Jason and I expect great things from him in the future. Emily Armbruster was also a dream collaborator in this period. Her data are unfailingly beautiful and I know I can always trust her analyses. I'm thankful to have had the opportunity to work together.

In the final period of my Ph.D., I started to work with Santiago Lopez, one of the most conscientious and kind people I know. One thing I admire about Santi is that he has no interest in the noise of popularity or competition. He is a scientist's scientist. Through Santi and his iconoclasm, I've refined my appreciation of nature - its sprawling complexity can be represented in elegant simplifications, but Santi helped me realize that much more nuance lies outside of our reductionist 'solutions'. Key to this period was also Claire O'Brien. Claire joined me as an already accomplished undergraduate, and she exceeded all expectations of what a student busy with classes and college life could do. I came to rely on Claire for experiments, analyses, and fun chats at the bench, and she was instrumental to the final part of this dissertation. I'm thrilled she is on the path to doctoral studies, but I'll miss working with her terribly.

The network of mentors by my side provided me key advice over these five years. Jill Banfield and I got to know each other in the second half of grad school. Our tea chats about life, academia, and big, crazy scientific ideas were always inspiring. I'm lucky to know her and call her a mentor. Michael Marletta, too, is a wonderful, often hilarious mentor who provided priceless advice all throughout this work. I'll never forget how he perfectly predicted the nuclease-protease story months before we acquired the data.

Danielle Dube, my undergraduate mentor at Bowdoin College, is the reason I'm at Berkeley. She believed in me back when I was struggling in her Biochemistry course, and she helped me realize I am capable of doing real research when I was at my least confident. I owe Danielle so much. Danielle and Karen Moulton continue to be constant sources of advice and support. Thanks also to Peter Seeberger, Eric Sletten, Conor Crawford, James Nuñez, Arwen Pearson, Jo Zaitsev-Doyle, Matt Francis, and Dave Savage for helping to propel me into and through grad school.

Kevin Wasko was my bay-mate for all of graduate school and became a good friend and confidante. He is an incredible scientist and, although thermodynamically implausible, I believe he could accomplish reverse translation if he set his mind to it. Noam Prywes was my mentor in a previous rotation, climbing buddy, and coffee partner. I'm thankful to him for introducing me to the works of Jacob, Monod, and Gould. Maria Lukarska and I have been friends since day one of my rotation. I know I can always rely on Maria for good book recommendations, symphony tickets, and a sympathetic ear. Maria, Hannah Nilsson, and Lou Devanneaux were my housemates/supper club companions. They are treasured friends with whom I can talk about anything for hours. Thanks also Amy Eggers, Brittney Thornton, Arushi Lahiri, Leo Song, Colin Robinson, Peter Yoon, Kenneth Loi, Ryan Kiracofe, Cullen Ridgway, Garyk Brixi, Lillian Petersen, Christian Gostout, Jacob Baskes, and many, many more people for being great colleagues and friends. You all made this time so wonderful. I must of course mention Kaihong Zhou and Jenny Ye, two of the hardest workers I've met. I loved chatting with Kaihong and Enbo, and I know no lab manager will ever approach Kaihong's greatness.

To Sophia Swartz I owe my heart and my world. Sophia and I met at a lab happy hour and quickly became inseparable. Our life together is a gift that's full of gifts - coffee left with a cute note on the lid, snacks stashed away in drawers, sunny 'touch-grass' sessions snuck in throughout the day, a nice little techno song, Monday gym and Tuesday breakfast and Friday Coro, deep conversations about science and the soul, tickets to concerts, wikipedia pages ranging from skinks to stramenopiles... Sophia is always there for me, whether it's during a difficult moment, to share a hard-won accomplishment, or to just enjoy each other's company. This work would not have been possible, nor this experience remotely as wonderful, without her. She likes to joke that tough days in lab make her "lose her sparkle," but I know that she actually does sparkle, and that her sparkle increases in luminosity every day. This Ph.D. was in many ways a setting for our relationship, and it pains me to leave its comfort. But I'm so excited for what's next, and I am certain we can do anything together.

Lastly, to my family. I owe everything to them. Mom, Dad, Andrew, Grandma, Abuela, Grammy Leni, Grandpa III, Aunt Karen, Uncle Rob, Chris, Steve, Aunt Merrell, Chloe... you never cease believing in me for a single second. I don't think I could ever really express in words what your love means to me. This is all for you. I love you all.

Chapter 1

Introduction

1.1 *Self and non-self*

How do organisms repel predators and harmful substances while protecting and nourishing themselves? These essential functions are provided by an overlapping set of pathways collectively referred to as the immune system. All cellular lifeforms are thought to use immune systems of some kind. To work appropriately, an organism's immune system must define the boundary between itself and its environment, and it must also distinguish between normal host components and the machinery of invaders like viruses should they infiltrate the inner sanctum of the cell. The task of immunity is therefore defining cellular identity, both externally and internally. The cell must distinguish *self* from *non-self* to survive.

The concept of immune *selfhood* and cell identity emerged in the nascent period of immunology with Metchnikoff¹ and found its formal articulation a century later with Janeway². Since then, the immune *self* has become an organizing principle guiding molecular immunology. Immune systems are now known to primarily function by recognition of biomolecules uniquely present in foreign entities like pathogens. These motifs are known as pathogen-associated molecular patterns, or PAMPs³. In humans and many other animals, for example, a subset of toll-like receptors (TLRs) bind specifically to lipopolysaccharide, a ubiquitous molecular pattern found in the membrane of gram negative bacteria⁴. RIG-I-like (Retinoic acid-inducible gene I-like) receptors (RLRs) are another class of sensory proteins that detects the accumulation of double-stranded ribonucleic acid (dsRNA), a common replication intermediate of viruses attempting to hijack the host cell⁵. In the case of TLRs, RLRs, and numerous other immune components, the PAMP is an invariant, highly evolutionarily conserved structure, maximizing the spectrum of immune reactivity. Upon detection of the pathogen-associated signal, a set of downstream functions are activated in a cascade of immunological signalling, culminating in responses ranging from recruitment of specialized cells to the site of infection to altruistic self-destruction to halt pathogenic replication^{6,7}. Invading entities respond in kind to host immunity, evolving variants of recognized structures and deploying their own repertoire of defenses to deactivate host immunity⁸. The back-and-forth between host and pathogen is commonly analogized as an 'arms race' or a 'struggle.' Detection of *non-self* drives the evolution of immune genes, organisms, and ecosystems.

1.2 Identity crisis

Several additional, recently discovered functions of immune systems prompt revision to the *self/non-self* model. Difficulties began when it was discovered that natural killer cells respond to missing *self* proteins and other stress signals in neighboring cells⁹, using solely *self* cues as direct signals for activation. Principles of immunity in response to changes in *self* have since been extended to encompass a wide range of immune reactions to so-called damage-associated molecular patterns, or DAMPs¹⁰. DAMPs are endogenous biomolecules in the host organism that are thought to release during stress, trauma, or infection and signal 'danger' to the host. Like PAMPs, dedicated pattern recognition receptors are thought to bind DAMPs to initiate a downstream response¹¹. One example of a DAMP from human biology is the cytokine IL-1 α . If IL-1 α is detected outside its 'normal' location in the nucleus, the cognate receptor IL-R1 activates a cascading inflammatory response to target the causative injury: skin epithelial cells produce this cytokine constitutively to provide a rapidly deployable 'alarm' in case the barrier to the *self* is breached¹². Despite some unresolved controversy over the precise nature of DAMP signalling¹³, the emergence of a *self*-only mode of sensing calls into question the very nature of immune discrimination.

One consequence of pattern-based discrimination is that *self* and *non-self* are not absolute categories but endpoints on a continuum of molecular similarity. Autoimmune responses manifest when the immune system mistakes *self* (or damaged *self*) for *non-self*, harming the fitness of the host organism¹⁴. For example, in cases of rheumatic fever, antibodies against cell surface antigens of the causative bacterial pathogen cross-react with structurally similar proteins in cardiac tissue, redirecting the immune response against the host's own heart¹⁵. In this case, the antigen is similar enough to *self* that the immune system cannot reliably distinguish between them. The very existence of autoimmunity reveals that *self/non-self* discrimination operates on thresholds tuned by evolution for population-level fitness, rather than the level of individuals¹⁶. Moreover, if *self/non-self* discrimination is imperfect even in complex vertebrate adaptive immune systems, it invites deeper questions about how simpler immune architectures, like those found in bacteria, handle the same problem.

What qualifies as *self* in a world we now appreciate as replete with symbiosis and commensalism? Within animals, for instance, reside a vast reservoir of microbes¹⁷, many of which provide beneficial functions for the host like nutrition¹⁸ and pathogen resistance¹⁹. Do microbes residing within animals constitute an extended form of *self*? Or are microbiome members merely tolerated in immune-privileged areas? Consider also that nearly all cells harbor mobile genetic elements, the 'alternative' class of evolving biological entities that includes viruses and transposons. In many cases, these 'invaders' help resist other infections and contribute to genetic diversity in their hosts²⁰. In fact, domesticated retroviral machinery is responsible for the evolution of the placenta in mammals²¹, and latent viruses are now thought to have contributed to

expansion of brain size in our hominid ancestors²². Clearly, immunology must adapt to an increasing ecological sensibility within the cell, within the organism, and beyond.

The immune *self* appears increasingly amorphous in an interconnected biological world. The *self/non-self* dichotomy is powerful, but uncritical adoption of its inherent metaphysical reductionism risks obscuring the nuances that animate complex and important immune interactions. In this work, I describe new mechanisms of immune sensing that deepen conceptions of immune *self*.

1.3 Ancestral immunity

This discussion has thus far considered immunity exclusively in the context of mammals, and even within that narrow frame major aspects that would be essential for a more comprehensive treatment have been omitted: the distinction between innate and adaptive immunity²³, clonal selection²⁴, and tolerance²⁵, to name a few. But the challenge of *self/non-self* discrimination is not unique to animals.

Immune identity must have emerged very early in the evolution of life. Immune genes are found in all cellular organisms, from eukaryotes, the organelle-bearing domain to which we humans belong, to prokaryotes, the vast superdomain of unicellular, organelle-free microbes encompassing archaea and bacteria like *Escherichia coli*. Immune genes comprise 7% of all human protein-coding genes²⁶, while in bacteria, this figure may lie well over 5% on average²⁷. Amazingly, despite billions of years of evolution separating our common ancestor, a significant fraction of immune systems in eukaryotes and prokaryotes are remarkably similar in structure and function^{28,29}. These parallels are too numerous, and the phylogenetic evidence too convincing, to attribute to evolutionary convergence alone³⁰. Rather, deep conservation of immune components reflects their outsize contribution to cellular fitness and the constant threat posed by infection throughout the history of life on Earth. That immune machinery has been so faithfully conserved across billions of years of evolution speaks to the absolute necessity of immunity for what François Jacob called “the dream of every cell: to become two cells”³¹.

The existence of *bona fide* prokaryotic immune systems, much less systems with homology to those of eukaryotes, is a recent development and area of active research. At the advent of molecular biology, hints of immune *self/non-self* discrimination mechanisms appeared in work on exclusion systems by Luria and Human³². These systems were later found to comprise the most abundant immune systems in all prokaryotes, restriction modification systems (RM), which function by methylating DNA motifs in the host genome to mark *self* and inhibit cognate nuclease that recognize the unmodified, *non-self* sequence³³. Later, CRISPR-Cas systems (clustered, regularly interspaced short palindromic repeats, CRISPR-associated) were found to function in an adaptive fashion^{34,35}, recording *non-self* sequences in a genomic ledger that guides nucleases to cleave matching sequences^{36–38} (discussed in greater detail in Chapter 2). Additional loci, like various abortive infection modules and toxin-antitoxin systems³⁹, were also found to protect against prokaryotic viruses. Still, these systems were largely understood as isolated, rather than components of a conserved immune system and logic.

The full scope of prokaryotic immunity only came into focus with the maturation of comparative genomics and sensitive computational methods for detecting remote homology. Pioneered by Koonin, Aravind, and colleagues⁴⁰ and later extended by Sorek⁴¹, Bernheim⁴², Gao⁴³, and others, this work revealed the depth and breadth of the

prokaryotic immune repertoire. For example, signaling NTPases of the STAND superfamily (signal transduction ATPases with numerous domains)^{44,45}, TIR (toll-interleukin receptor) domains⁴⁶, and cGAS-like (cyclic GMP-AMP synthase-like) receptors are also encoded in prokaryotic genomes^{47,48}. Subsequent molecular characterization revealed that these shared domains function similarly despite immense phylogenetic distance from their eukaryotic homologs. That such components are found not only across the eukaryotic kingdoms but also archaea and bacteria points to an ancient immune system that predates the divergence of the major domains of life. Indeed, ancestral reconstructions of the genome content of LUCA, the last universal common ancestor of all extant cellular lifeforms, suggest the presence of at least a CRISPR-Cas system and likely many other immune functions⁴⁹. Over the past five years, the emergence of *de novo* structural prediction and comparison methods have deepened this realization^{50,51}, revealing homologies too distant to detect by sequence alone. These methods enable reconstruction of how immune mechanisms have been inherited, shuffled, and repurposed across billions of years of evolution.

1.4 Bacteriophages

Bacteriophages (phages) are bacterial viruses that form the driving force behind the evolution of prokaryotic immune systems¹. Phages are by far the most abundant biological entities on Earth. An estimated 10^{31} phages populate the biosphere, outnumbering their 10^{30} bacterial hosts 10:1⁵². In the case of lytic phages that kill their hosts, 10s or 100s of phage progeny can be produced from a single initial virion. Consequently, phage infection is a major cause of bacterial mortality: in marine ecosystems, up to 40% of bacteria are thought to succumb to phage infection every day⁵³. Phage may account for as much as 15% of the daily turnover of biomass on Earth⁵⁴. Considering their lethality, the ecology of phage-host interactions is more nuanced than binary predator-prey dynamics. In the human gut, for example, lytic phages are rare⁵⁵. Most phages persist as dormant prophages integrated into bacterial genomes, excising and entering the lytic cycle only under specific conditions. Prophages are found in nearly every bacterial genome, yet their biology remains understudied relative to their lytic counterparts⁵⁶. This imbalance likely obscures much of the true landscape of phage-host coevolution.

1.5 Pan-immunity

Despite deep ancestry of immune systems between prokaryotes and eukaryotes, there are several differentiating features of prokaryotic immune systems that merit special consideration. Most striking is the density and diversity of defense systems encoded within a bacterial and archaeal single genome. A typical bacterium may harbor restriction-modification systems, CRISPR-Cas loci, abortive infection modules, and numerous other defenses simultaneously⁴². This redundancy must have evolved in light of an adaptive advantage. Individual defense systems are often effective against only a subset of phages, and phages in turn evolve counter-defenses that can neutralize any single line of defense. Thus, a broad immune arsenal hedges against the functional and genetic diversity of the viral world⁵⁷.

Modularity is a dominant evolutionary mechanism driving the diversification of antiviral immune systems in prokaryotes - despite their astonishing diversity, systems draw from a relatively small pool of conserved protein domains^{43,46,58-60}. These conserved domains often encode different effector mechanisms and exchange rapidly to enable rapid diversification of immune loci^{28,61}. Much remains unknown about the emerging principle of modularity in immune evolution: how individual modules function across different genetic contexts, what consequences their exchange has for immune activity, and why certain modules recur in certain systems and not others. In the final chapter of this work, I make a case for using modularity as a means of linking otherwise divergent immune systems through the shared mechanisms of their component parts.

¹Archaea are technically infected by archaeal viruses. Archaeal viruses are equated to phages here for simplicity.

As in the eukaryotic world, reactive prokaryotic immune systems carry risks. Defense systems are known to activate in response to host-encoded, *self* cues. CRISPR-Cas systems, for example, occasionally acquire spacers from *self* rather than from invading mobile genetic elements, licensing cleavage of the host genome⁶². Abortive infection systems, which use programmed cell death or stasis to curb infection, kill or arrest the host cell. Expression of these systems is protective against phages, but carries the high cost of spurious activation⁶³. One consequence of the tradeoff of carrying a large immune arsenal is that systems are frequently lost from microbial genomes over short evolutionary timescales⁶⁴. In fact, immune operons are some of the most dynamic genes after mobile genetic elements themselves, contributing up to 90% of intraspecies variation in *Vibrio cholerae*, for instance⁶⁵.

Rapid immune turnover is balanced by acquisition. Antiphage immune systems spread through horizontal gene transfer (HGT); they are often found on plasmids, transposons, and phages themselves⁶⁶. The tendency for prokaryotic immune loci to cluster in genomic 'defense islands' enriched with mobile elements is the basis for the recent acceleration in their identification⁴¹. The functional consequence of mobilizeable immune hotspots is relevant at the population-level. No single cell encodes all possible defenses, but the 'pan-genome' of a bacterial population, including different species compatible for HGT, collectively harbors a vast, accessible reservoir of immune systems⁶⁷. This 'distributed diversity' ensures that at least some members of the population carry an effective defense, and that survivors may propagate fit immune systems relevant to the local phage context. Immunity in prokaryotes is thus an emergent feature of populations shaped by the interplay of selection, horizontal transfer, and relentless coevolutionary pressure.

1.5 This dissertation

In this dissertation, I describe three studies that investigate how prokaryotic immune systems sense threats, discriminate *self* from *non-self*, and diversify through modular exchange of functional components.

Faithful acquisition of new immune memories is essential for CRISPR-Cas adaptive immunity, yet the mechanisms ensuring *self/non-self* discrimination during this process remain incompletely understood. In Chapter 2, I present cryo-EM structures and biochemical analyses of a CRISPR trimmer-integrase, a natural Cas1:Cas2-exonuclease fusion that coordinates DNA capture, trimming, and integration. These analyses reveal how an immunological tag called the protospacer adjacent motif (PAM) is retained and the spacer processed. The PAM sequence is cleaved prior to genome integration, marking spacers as *self* and preventing autoimmune targeting, representing a new, DNA topological mechanism for *self/non-self* sensing.

Chapter 3 turns to Hachiman, a widespread antiphage defense system whose mechanism was previously unknown. I show that Hachiman is a nuclease-helicase complex in which the sensor helicase unleashes cell killing upon detection of infection-associated *self* DNA damage, an echo of the DAMP-sensing paradigm discussed above. The Hachiman helicase is also suggested as a putative ancestor of an important family of helicases in eukaryotes.

In Chapter 4, I describe the mechanism of nuclease-protease gene pairs, an effector module, that has been repeatedly and independently exchanged between otherwise unrelated antiphage immune systems. I show that the nuclease is a proenzyme activated by its partner protease. Then, I show how modular inheritance enables the discovery of new immune loci, an approach I term “guilt-by-modularity”. This strategy led to the discovery of caspase-nuclease, or *canu*, immune systems with functional parallels to caspase-mediated immunity in eukaryotes.

Together, these studies illuminate how prokaryotic immune systems solve problems of *self/non-self* discrimination, autoimmunity, and diversification. Solutions to these problems are often shared across the domains of life.

Chapter 2

Genome expansion by a CRISPR trimmer-integrase

This chapter is adapted from: Wang, J.Y.*, Tuck, O.T.*, Skopintsev, P.*, Soczek, K., Li, G., Al-Shayeb, B., Zhou, J., and Doudna, J. (2023). Genome expansion by a CRISPR trimmer-integrase. *Nature* 618, 855–861.

* these authors contributed equally

J.Y.W. conceived the project and designed experiments with O.T.T., P.S., G.L. and J.A.D. J.Y.W., O.T.T., G.L. and J.Z. performed protein purification and in vitro biochemistry. P.S. performed cryo-EM data collection with K.M.S. and O.T.T. P.S., O.T.T. and K.M.S. generated cryo-EM maps and built atomic models with input from J.Y.W. B.A.-S. identified the CRISPR–Cas locus of interest and performed bioinformatics experiments. O.T.T., J.Y.W. and J.A.D. wrote the manuscript with input from all of the authors.

2.1 Abstract

CRISPR-Cas adaptive immune systems capture DNA fragments from invading mobile genetic elements and integrate them into the host genome to provide a template for RNA-guided immunity³⁶. CRISPR systems maintain genome integrity and avoid autoimmunity by distinguishing between *self* and *non-self*, a process for which the CRISPR-Cas1:Cas2 integrase is necessary but not sufficient^{68–71}. In some microbes, the Cas4 endonuclease assists CRISPR adaptation^{72,73}, but many CRISPR-Cas systems lack Cas4⁷⁴. We show here that an elegant alternative pathway in a Type I-E system employs an internal DnaQ-like exonuclease (DEDDh) to select and process DNA for integration using the protospacer adjacent motif (PAM). The natural Cas1:Cas2-exonuclease fusion (trimmer-integrase) catalyzes coordinated DNA capture, trimming and integration. Five cryo-EM structures of the CRISPR trimmer-integrase, visualized both before and during DNA integration, show how asymmetric processing generates size-defined, PAM-containing substrates. Before genome integration, the PAM sequence is released by Cas1 and cleaved by the exonuclease, marking inserted DNA as *self* and preventing aberrant CRISPR targeting of the host. Together, these data support a model in which CRISPR systems lacking Cas4 use fused or recruited^{75,76} exonucleases for faithful acquisition of new CRISPR immune sequences.

2.2 Introduction

Prokaryotes use CRISPR-Cas (clustered regularly interspaced short palindromic repeats, CRISPR associated) adaptive immune systems to create a sequential genetic record of infection⁷⁷. Transcription and processing of CRISPR sequence arrays, which consist of short repeats and ~30 base pair (bp) foreign DNA-derived spacers^{36,68,78,79}, yields mature CRISPR RNAs (crRNAs) that guide interference of matching genetic material, protecting the host against recorded sequences^{37,80–84}.

The Cas1:Cas2 integrase drives CRISPR array evolution by selecting and inserting new spacers^{69,85,86}. Cas1₄:Cas2₂ is a heterohexameric complex that specifically recognizes DNA fragments (protospacers) containing a ~30 bp segment with short single-stranded 3' overhangs^{35,86,87}. In DNA-targeting CRISPR systems, protospacer selection requires a flanking 2–5 bp sequence known as the protospacer adjacent motif (PAM), which is a key component used to distinguish *self* from *non-self* and evade autoimmunity. The PAM is selected during DNA capture, but is removed before host genome integration. Coordinated selection and removal of the PAM ensures Cas interference modules target true invasive elements instead of the host CRISPR array^{71,88,89}.

Diverse mechanisms of PAM selection and removal underscore the importance of the PAM for maintaining both adaptive immunity and genome integrity during CRISPR sequence acquisition^{74–76,90–92}. In CRISPR systems including type II-B, some type V, and type I-A, -B, -C, -D, and -G, the Cas4 endonuclease performs PAM selection and

processing^{72,73,90,93,94}. However, ~40% of CRISPR subtypes lack Cas4⁷⁴. In Cas4-less systems such as the type I-E system in the common laboratory *E. coli* K12 strain, Cas1 contains a PAM-binding pocket which is believed to play a role in protospacer precursor (prespacer) selection^{75,87}. However, whether Cas1 cleaves the PAM in a similar manner to Cas4 or relies on host nucleases to perform this function remains unclear^{75,87}. Recent in vitro studies identified host exonucleases capable of aiding Cas1:Cas2 in prespacer substrate trimming^{75,76,95}. Standalone exonucleases such as the DnaQ-like exonuclease class DEDDh are widespread ancillary components present in every CRISPR-Cas type⁷⁴. There are also Type I-E systems containing a natural Cas2-DEDDh exonuclease fusion^{74,92,96}, further implying a functional link between exonucleases and the CRISPR integrase. These systems provide a model for studying coordination between host exonucleases and the CRISPR integrase.

Here, we reconstitute CRISPR sequence capture, processing, and integration by a naturally occurring *Megasphaera* NM10-related Cas1:Cas2-DEDDh complex. We show Cas1:Cas2-DEDDh preserves the PAM during prespacer processing and the first step of integration. The PAM is removed before completing full integration. The DEDDh active site, rather than Cas1⁸⁷, is responsible for both initial 3' overhang trimming and PAM removal. This mechanism is distinct from that of Cas4, which cleaves the PAM endonucleolytically, suggesting a divergent role for host exonucleases in PAM processing⁷⁵. The integrase regulates DEDDh exonuclease activity by a ruler-guided, 'gatekeeping' mechanism that coordinates processing and defines the length of integrated DNA. Cryo-electron microscopy (cryo-EM) structures of Cas1:Cas2-DEDDh bound to prespacer DNA with or without the PAM show how Cas1:Cas2 recognizes the sequence and protects it from DEDDh-mediated trimming. Conformational analysis of half-integration structures suggests that, once anchored into the CRISPR array, DNA bending engages the C-terminal region of Cas1, which in turn exposes the PAM for removal, enabling full integration. Our findings provide a general mechanism for exonuclease-assisted PAM processing and demonstrate that CRISPR systems evolved diverse mechanisms to ensure robust immunity against parasitic elements and avoid autoimmunity.

2.3 Results

2.3.1 Cas1:Cas2-DEDDh substrate generation

CRISPR adaptation relies on the recognition, capture and processing of suitable DNA integration substrates from foreign sources (**Fig. 2-1A**). These integration substrates (prespacers) require nucleolytic processing to generate fragments of uniform length. To investigate the predicted exonuclease domain of Cas2-DEDDh, we expressed and purified Cas1 and Cas2-DEDDh from a type I-E *Megasphaera* NM10-related CRISPR system and tested DNA substrate processing in vitro (**Fig. 2-1B**). The size of spacers in the *Megasphaera* CRISPR array and preferences of the related I-E *E. coli* Cas1:Cas2 integrase suggest the preferred substrate is a 23-bp DNA duplex with 5-nt

single-stranded 3' overhangs^{35,87}. To test DNA processing, we assayed Cas1 and Cas2-DEDDh trimming activity using 5' fluorophore-labeled prespacer substrates containing a 23-bp duplex region and extended single-stranded 3' overhangs of varying lengths. Cas1 and Cas2-DEDDh each exhibit nuclease activity in isolation, yielding distinct products without apparent functional relevance (corresponding with partial cleavage or complete trimming of the 3' end, respectively). Only the reconstituted Cas1:Cas2-DEDDh complex generates substrates equivalent in size to spacers in the host CRISPR array (**Fig. 2-1B**). Variation of substrate sizes showed the integrase requires a 23-bp duplex for functional processing (**Supplementary Fig. 2-1A, B**). We next tested whether the DEDDh active site is responsible for processing activity by using a catalytically inactive DEDDh mutant (D132A). DNA cleavage assays indicated that the mutant complex does not process prespacers (**Fig. 2-1C**). Taken together, these data demonstrate that the complete Cas1:Cas2-DEDDh complex is necessary for prespacer processing and that the DEDDh active site provides the requisite nucleolytic activity.

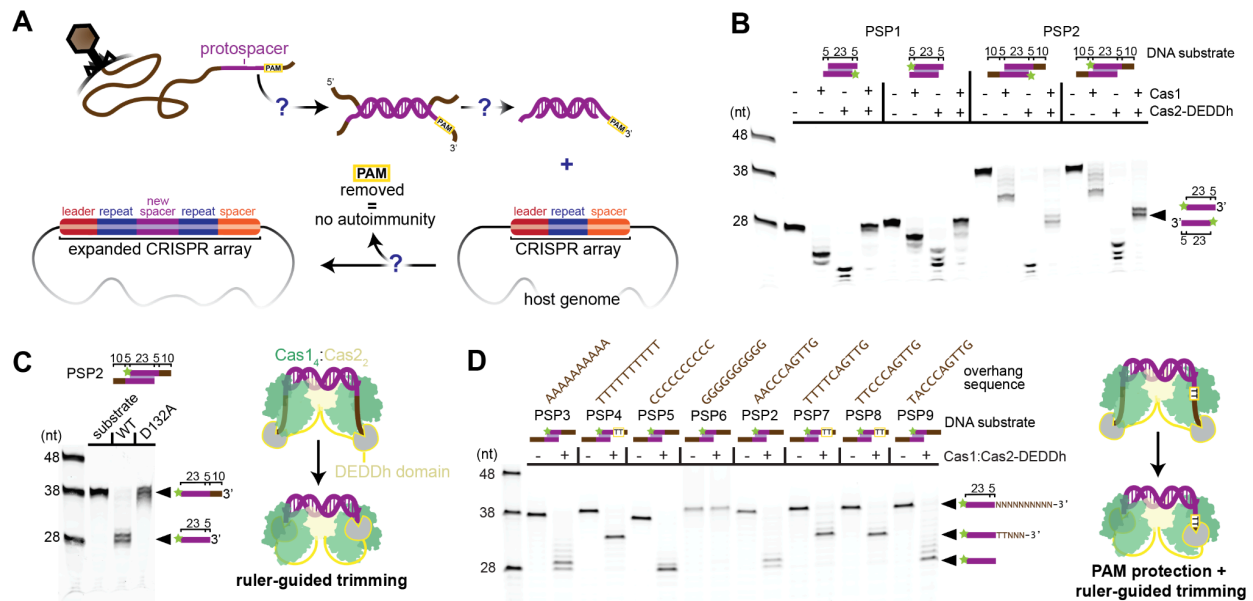


Figure 2-1. Cas1-Cas2/DEDDh processes prespacers to the correct size for integration and protects a TT PAM. **A)** Open questions in CRISPR adaptation. **B)** Processing of fluorescently labelled prespacer substrates with a 23 bp duplex and different overhang lengths by Cas1:Cas2-DEDDh down to 5–6-nucleotide (nt) single-stranded 3' overhangs. Prespacer 1 (PSP1), 23 bp duplex with 5-nucleotide 3' overhangs; prespacer 2, 23 bp duplex with 15-nucleotide 3' overhangs. The star indicates the 6-carboxyfluorescein label. **C)** WT and mutant Cas1:Cas2-DEDDh (DEDDh(D132A)) prespacer processing. **D)** Processing of substrates with variable 3' overhangs and the model of observed PAM protection and ruler-guided trimming by the DEDDh domain.

Time-course assays suggest similar processing efficiencies for prespacer substrates with varying overhang lengths (**Supplementary Fig. 2-2A-D**). Fluorescently labeled prespacers were incorporated into an integration target plasmid (pCRISPR) containing

a shortened version of the natural *Megasphaera* CRISPR array. Kinetic analysis implies higher relative integration efficiency with the canonical substrate (23-bp duplex with 5-nt single-stranded 3' overhangs) compared to prespacers with extended overhangs. Reaction with the canonical substrate generated ligation products after 2 min, while prespacers with extended overhangs required 10 min for detection. Thus, Cas1:Cas2-DEDDh provides a molecular ruler against which DEDDh trims prespacers.

To determine the effect of the PAM, we varied the overhanging region, generating a small prespacer library against which the PAM could be inferred (**Fig. 2-1D**). We determined that Cas1:Cas2-DEDDh recognizes 5'-TT in the PAM position. In the absence of a TT PAM, DEDDh trims prespacer strands to the integration-competent size (28 nt). However, the presence of a TT PAM in the correct position (nucleotide positions 29 and 30 relative to the 5' end) results in partial trimming of the PAM-containing strand, precisely 3 nt away from the PAM, yielding a 33 nt product. We hypothesized that partial trimming was the result of sequestration by a PAM binding pocket in Cas1⁸⁷.

2.3.2 Cas1 binds the PAM and blocks DEDDh trimming during prespacer processing

We next sought to elucidate the structural basis for prespacer processing and PAM protection. Cryo-EM was used to solve 3.1 Å and 2.9 Å resolution structures of Cas1:Cas2-DEDDh complexed with prespacer substrates with and without a TT PAM, respectively (**Fig. 2-2A-D**). The *Megasphaera* Cas1:Cas2 retains the canonical heterohexameric architecture⁹⁷, with two Cas1 dimers (denoted Cas1 and Cas1', a or b subunit) bridged by a central Cas2 dimer (**Fig. 2-2A**). A 23 bp duplex sits atop the complex, while 5 nt single stranded 3' overhangs extend into the clefts formed at opposite Cas1 interfaces. The DEDDh domain could not be resolved in the PAM-absent dataset, presumably due to a high degree of conformational flexibility conferred by the 38 amino acid linker between DEDDh and Cas2, and because all 3' ends are buried within the complex, protected from exonuclease activity (**Fig. 2-2C**). In the PAM-containing dataset, however, the DEDDh domain was resolved by iterative classification and three-dimensional refinement (see **Supplementary Fig. 4** and the Methods section).

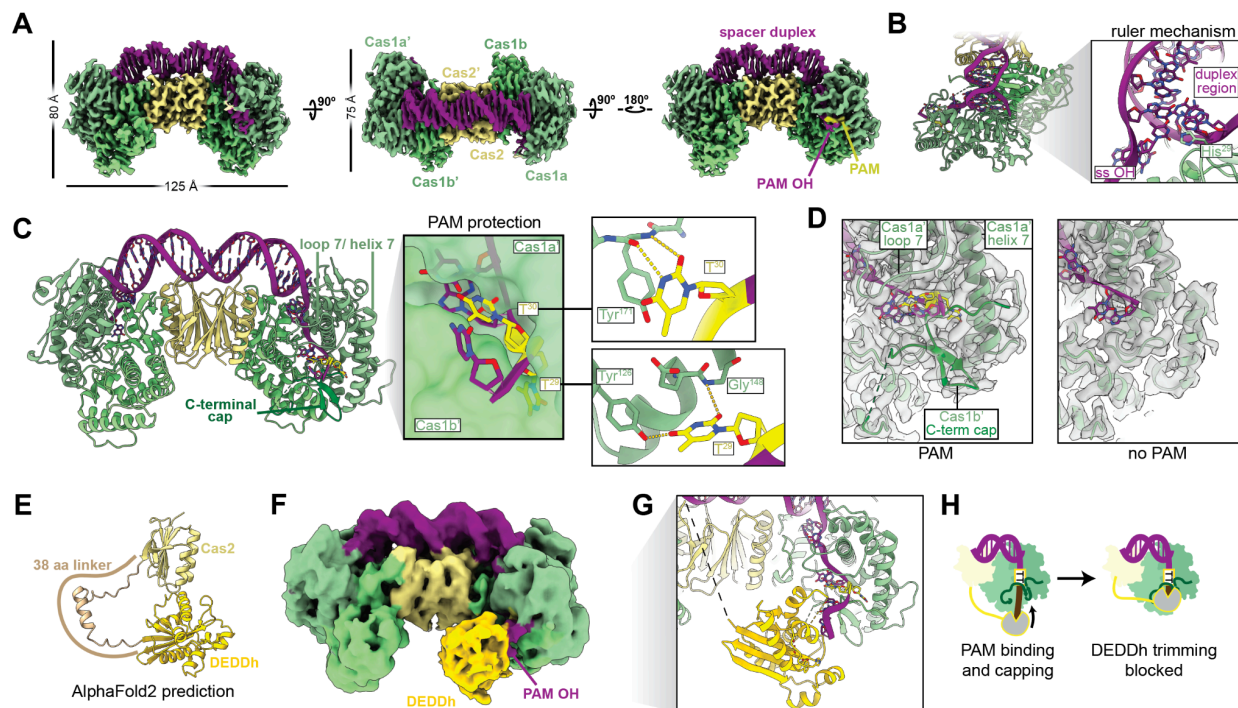


Figure 2-2. Molecular detail of Cas1–Cas2/DEDDh during prespacer processing. **A)** Orthogonal views of the final cryo-EM densities for Cas1:Cas2-DEDDh bound to a prespacer containing a phosphorothioated TT PAM (threshold, 0.200). **B)** The structure of PAM-bound Cas1:Cas2-DEDDh, depicting one of two His29 residues dictating duplex length. ss OH, single-stranded overhang. **C)** The structure viewed from the PAM side (left). Middle, surface depiction of the cleft between Cas1a' and Cas1b'. Right, sequence-specific contacts made with each PAM thymine. **D)** Comparison of PAM and non-PAM densities with atomic models overlaid at a threshold of 0.200. Term., terminal. **E)** AlphaFold 2 prediction of the structure of Cas2-DEDDh. aa, amino acid. **F)** Side view of unsharpened cryo-EM density in which DEDDh was resolved (threshold, 0.033). **G)** Hybrid structure containing the DEDDh domain with detail at the PAM–DEDDh interface, with catalytic DEDDh residues shown. The black dashed line represents the unstructured linker between Cas2 and DEDDh domains. **H)** Model for PAM protection.

Cas1:Cas2-DEDDh dictates prespacer duplex length with an internal ruler that ensures spacers are equivalent in length (**Fig. 2-1B, C**). Dual Cas1 His29 histidyl residues measure out a 23-bp DNA duplex by π -stacking with the terminal base pairs of the double stranded region, clasping the prespacer strands and marking the start of the single stranded 3' overhang (**Fig. 2-2B**). Biochemical processing assays demonstrate that only the 23 bp duplex is both tolerated and trimmed to the integration-competent length (**Supplementary Fig. 2-1A, B**).

Processing experiments suggest the PAM is protected initially from DEDDh-mediated prespacer processing (**Fig. 2-1B-D**). To understand how the integrase sequesters the PAM, we performed cryo-EM analysis on the integrase complex bound to a PAM-containing prespacer with a phosphorothioate backbone modification at the predicted site of DEDDh activity, with the intention of stalling processing.

Sequence-specific interactions with loop 7 (N162 to D179) and helix 7 (M139 to Y161) in the Cas1a' subunit in the resultant density rationalize PAM recognition (**Fig. 2-2C**). The first PAM thymine is buried in a pocket in Cas1a', where hydrogen bonds formed with Tyr126 and Gly148 may enhance binding affinity (**Fig. 2-2C**, right-top). The second PAM thymine π -stacks with Tyr171, which positions T30 to hydrogen bond with the Tyr171 backbone amide nitrogen (**Fig. 2-2C**, right-bottom). When the PAM is absent, substrates are fully trimmed, underscoring the necessity of sequence-specific interactions for asymmetric trimming and PAM protection (Fig 2-1D). Moreover, a beta hairpin in the C-terminal region of Cas1b' 'caps' sequestered nucleotides. Loop 7, helix 7, and the C-terminal cap are absent in the PAM-deficient density, suggesting these structural motifs participate in PAM protection (**Fig. 2-2D**).

The DEDDh domain was not visible in the initial PAM-containing structure, raising the question of how the integrase engages the DNA for ruler-guided trimming of sequestered nucleotides. To resolve the DEDDh domain, we iteratively classified and refined PAM-containing particles and found a density corresponding to DEDDh in a small subset of the total ensemble (**Fig. 2-2F**, **Supplementary Fig. 2-4**). Key features include a large protrusion only visible on the PAM side of the complex and an extended density attributable to additional phosphorothioate-containing nucleotides 3' to the PAM (**Fig. 2-2F**). Because the protuberance was low resolution, the model predicted by AlphaFold 2 for the DEDDh domain was docked into the density (**Figs. 2-2F, G**). The resulting hybrid model illustrates dynamics of DEDDh trimming and PAM protection. Catalytic DEDDh residues are poised to exonucleolytically cleave the overhang, but the PAM binding pocket and the C-terminal loop occlude DEDDh procession, blocking cleavage of the PAM and 2-3 additional nucleotides (**Fig. 2-2H**). Despite the high local concentration of nonspecific exonuclease relative to the substrate, this process is precise, in concordance with biochemical evidence (**Figs. 2-1B; 2-2G**). A natural consequence of protection is that the PAM must be cleaved downstream.

2.3.3 DEDDh trims the PAM after DNA half-integration

Although there is evidence for PAM protection during prespacer processing, the PAM must be removed prior to insertion into the CRISPR array to avoid autoimmunity. We analyzed Cas1:Cas2-DEDDh processing of DNA substrates designed to mimic intermediates of integration into the CRISPR array to determine the mechanism of PAM removal and resolve dynamics of the complex at the integration target site^{73,75}. Two substrates that mimic probable half-integration intermediates - the pre-PAM processing intermediate and the post PAM-processing intermediate (**Fig. 2-3A**, half-site substrates 1 and 2, respectively) - were synthesized and assayed in reactions with wildtype and catalytically-inactivated DEDDh complexes. Reaction with half-site substrate 1, which contains the unprocessed PAM, resulted in a 100 nt band corresponding to full-site integration (**Fig. 2-3A**). The same band was absent when DEDDh was catalytically inactive. Reactions with PAM-deficient half-site substrate 2 yielded the 100 nt full-site

integration product for both the WT and dead DEDDh complexes. These data suggest that, in the wildtype reaction, the PAM is fully removed prior to full-site integration. The lower intensity of the full-site integration product generated from half-site substrate 1 compared to that of half-site substrate 2 may be a result of inefficient PAM removal, also observed in kinetics assays (**Supplementary Fig. 2-2A, B**). The absence of the integrated product strand in the catalytically-inactivated DEDDh condition suggests the DEDDh active site executes PAM processing. Notably, PAM processing is necessary for full-site integration, and the integrase generates a precisely defined insertion product size. Hence, nonspecific exonuclease activity generates a ladder of ssDNA overhang fragments in the PAM-containing substrate strand, but only one of these fragment sizes is compatible with full-site integration. This single-nucleotide precision is a result of the Cas1:Cas2 ruler, which simultaneously defines the spacer size and acts as a gate that prevents PAM insertion into the CRISPR array (**Fig. 2-3A**). Since Cas1-mediated PAM protection was observed during prespacer processing, it is reasonable to assume that Cas1 releases the PAM for DEDDh trimming while engaged on the CRISPR array.

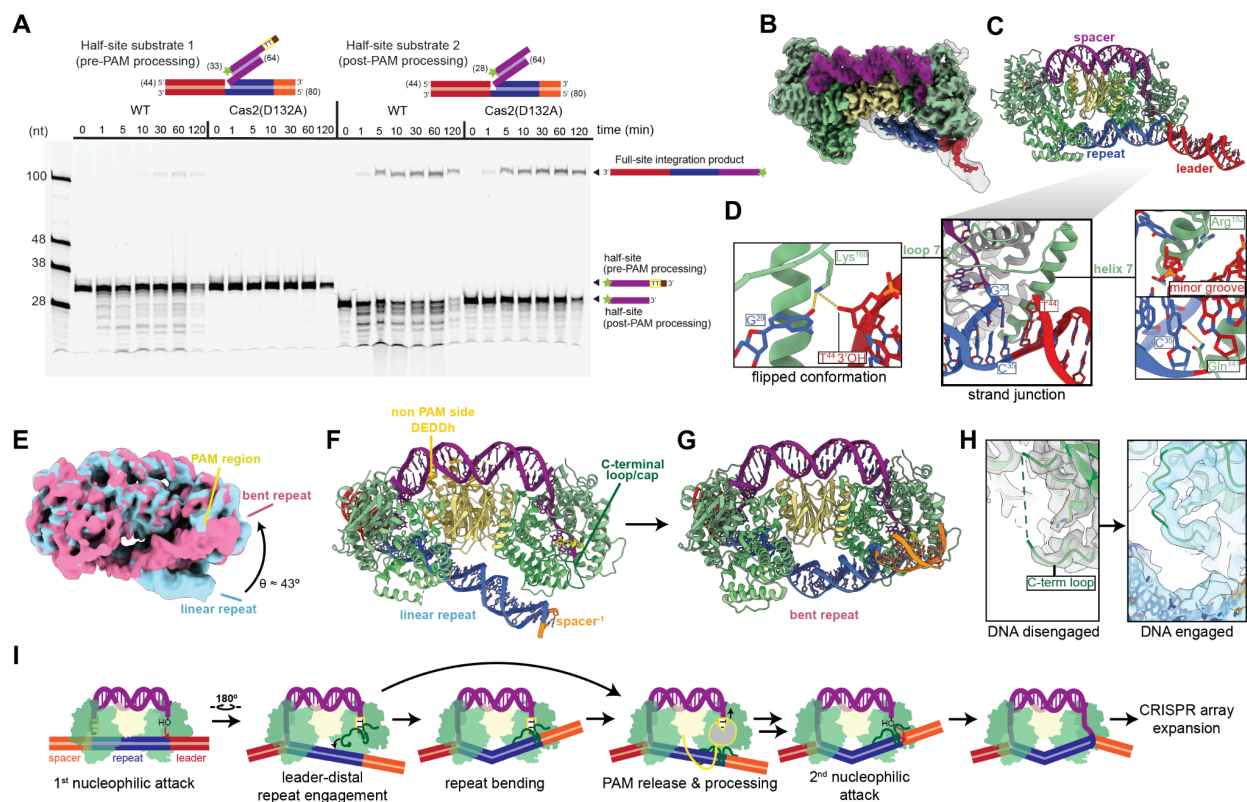


Figure 2-3. Biochemical and structural analysis of Cas1-Cas2/DEDDh PAM processing. **A)** Integration reactions by WT and mutant Cas1:Cas2-DEDDh (DEDDh(D132A)) half-site integration intermediates. **B)** Unsharpened (grey transparent; threshold, 0.05) and sharp (colour; threshold, 0.19) cryo-EM densities for Cas1:Cas2-DEDDh bound to the PAM-phosphorothioated half-site integration analogue. **C)** The structure of the initial half-site complex. **D)** Details of the junction of leader, repeat and spacer (middle). Left, a second, flipped conformation of G29. Right, helix 7 interactions. **E)** Unsharpened

maps of two repeat DNA conformations, linear (blue) and bent (red), both with a threshold of 0.05. **F)** Structure with linear, extended repeat DNA. **G)** Structure of bent repeat DNA. **H)** Comparison of C-terminal loop ordering in the prespacer (left) and linear repeat (right) structures. **I)** Model for PAM gatekeeping facilitated by the C-terminal loop/cap.

Evidence for DEDDh involvement in both prespacer processing and PAM cleavage led us to wonder which molecular cues prompt Cas1 to relinquish the PAM for digestion. Aiming to visualize DEDDh trimming and conformational changes in Cas1, we used cryo-EM to characterize Cas1:Cas2-DEDDh in complex with a DNA half-site analog containing phosphorothioate linkages at the PAM positions (**Fig. 2-3B**, **Supplementary Fig. 2-5A-C**). Neither the initial 3.1 Å density nor any heterogeneous states detected during cryo-EM data processing had density corresponding to the DEDDh domain on the PAM side of the complex (**Fig. 2-3B**). However, DEDDh was observed on the non-PAM side during DNA conformational analysis. We speculate that, in agreement with biochemical data, only the DEDDh domain can trim the PAM, and PAM trimming activity is required for full integration (**Fig. 2-3A**). Furthermore, the DEDDh PAM-trimming state may be transient. Phosphorothioate modifications only partially protected the PAM, which may disfavor resolution of the active DEDDh domain at the half-site (**Supplementary Fig. 2-6**).

The 3.1 Å half-site structure reveals previously uncharacterized interactions at the first integration strand junction (**Fig. 2-3B-D**). After the initial nucleophilic attack, Cas1:Cas2 induces bending of the leader-repeat target DNA⁹⁸. This deformation originates at the nick site, positioning repeat DNA between the leader-distal and -proximal Cas1 active sites. Inspection of the strand junction (**Fig. 2-3D**, middle) reveals interactions with the first base pair of the CRISPR repeat dictated by loop 7 and helix 7 of Cas1a'. The spacer-ligated first CRISPR repeat guanine G29 was found in two approximately equivalent conformations. In the first conformation, G29 forms a canonical base pair with C35 (**Fig. 2-3D**, middle). In the second conformation, G29 'flips' upward, making a specific contact with loop 7 lysine Lys168. The lysine also contacts the 3' hydroxyl of the leader (**Fig. 2-3D**, left). Gln141, which sits at the base of helix 7, also makes a nucleobase-specific contact with C35, the first bottom repeat nucleotide. Helix 7 is well positioned to insert into the minor groove of the leader DNA, but no nucleobase-specific contacts were obvious (**Fig. 2-3D**, right). Specific contacts with CRISPR array nucleotides likely play a functional role in targeting, as previously observed^{85,86,99}.

The 3.1 Å half-site complex reconstruction contains density corresponding only to the leader-proximal region of the CRISPR repeat (**Fig. 2-3B, C**). To probe dynamics at the leader-distal region, where PAM processing and subsequent full integration occur, we performed three-dimensional variability analysis (3DVA) on the particle set (**Supplementary Fig. 2-5A,D,E**)¹⁰⁰. 3DVA revealed heterogeneity in the location of the repeat/spacer end, with the CRISPR repeat DNA oscillating between linear and bent conformations (**Fig. 2-3E**). The DEDDh domain was visible only in the linear

conformation. Isolation and refinement of particle clusters representing maxima of the reaction coordinate gave linear and bent reconstructions at 4.1 Å and 3.9 Å resolutions, respectively (**Fig. 2-3E-G**).

In the linear structure, a Cas1b' C-terminal loop (L279 to S293) rich in charged residues is positioned near the major groove adjacent to the second integration target site (**Fig. 2-3F**). The corresponding density is absent in the 3.1 Å half-site and 2.9 Å prespacer-bound structures, indicating this loop participates in engagement with the CRISPR repeat on the PAM side (**Figs. 2-2C; 2-3G,H**). A C-terminal cap, which follows the C-terminal loop, protects the PAM and adjacent nucleotides from trimming by DEDDh (**Fig. 2-2H**). Intriguingly, the DEDDh domain was visible in the linear structure, but on the non-PAM side of the complex, where no overhang trimming occurs (**Fig. 2-3F**). The exonuclease sits in a cavity formed by the interface of the CRISPR repeat DNA, Cas1a/b', and Cas2, where it appears to contact the repeat DNA backbone and the N-terminus of the Cas1b (**Fig. 2-2F, Supplementary Fig. 2-7**). These interactions may stabilize beneficially constrain bending of the second integration target site or prevent trans activity.

The bent structure features a pronounced kink in the center of the repeat region (**Fig. 2-3G**). Disruption of a single A-T base pair and DNA unwinding appear to accommodate the strain induced by this pitch, though this assignment was made with low confidence due to the low local resolution at the bending site. Bending in the center of the CRISPR repeat symmetrizes the integration complex and draws the repeat/spacer junction towards the Cas1a' active site. Although PAM nucleotides are still present, likely due to their cleavage-blocking phosphorothioate modifications, the overhang nucleotides typically sequestered during ruler guided trimming are absent, suggesting that 3' trimming occurs in an intermediate step between linear and bent states. Only the first three nucleotides under the C-terminal loop of Cas1b' could be assigned, and the C-terminal cap density was largely unstructured. These observations, combined with the proximity of the second integration target site in the bent structure, suggest that engagement of the CRISPR DNA by the C-terminal loop induces uncapping of previously sequestered PAM nucleotides, followed by DEDDh- or host exonuclease-mediated trimming of the exposed 3' end to the ruler-defined length. Once the PAM is trimmed, full integration occurs (**Fig. 2-3A**). Structural and biochemical analyses imply a general mechanism of sequential PAM protection and cleavage, or PAM gatekeeping, which ensures that PAM-deficient protospacers integrated into the CRISPR array are marked as *self* and are equal in size (**Fig. 2-3I**).

2.3.4 Reconstitution of full CRISPR array integration

To determine how PAM sequence recognition and gated removal ensures accurate DNA integration, we reconstituted CRISPR substrate integration *in vitro*. An unprocessed prespacer (23-bp duplex with 15 nt single-stranded overhangs) containing the TT PAM was combined with Cas1:Cas2-DEDDh and pCRISPR.

Prespacer substrates encoded *BsaI* restriction sites in the duplex region to enable insertion of a chloramphenicol resistance gene. Upon transformation, only pCRISPR with full prespacer integration confers survival in a double selection assay (**Fig. 2-4A**)¹⁰¹. Fully integrated sequences provide additional evidence that Cas1:Cas2-DEDDh completes both prespacer processing and integration into the CRISPR array (**Fig. 2-4B**). The complex is specific for the CRISPR array and all integration events occur at repeat borders (**Supplementary Fig. 2-8B**). However, integration occurs at all three repeats present in the array, without specificity for the leader-proximal repeat, as seen in many CRISPR systems *in vivo*^{36,102} and consistent with structural data (**Fig. 2-3D**). Excess 3' overhangs are trimmed to within 1-2 nt of the expected length and the PAM is absent in all integrated sequences (**Fig. 2-4B**), in agreement with evidence at the level of the half-site (**Fig. 2-3A,I**). Reconstitution experiments provide complementary evidence for an alternative mechanism for PAM processing compared to Cas4, which uses a sequence-specific mechanism to cleave the PAM endonucleolytically^{73,93}.

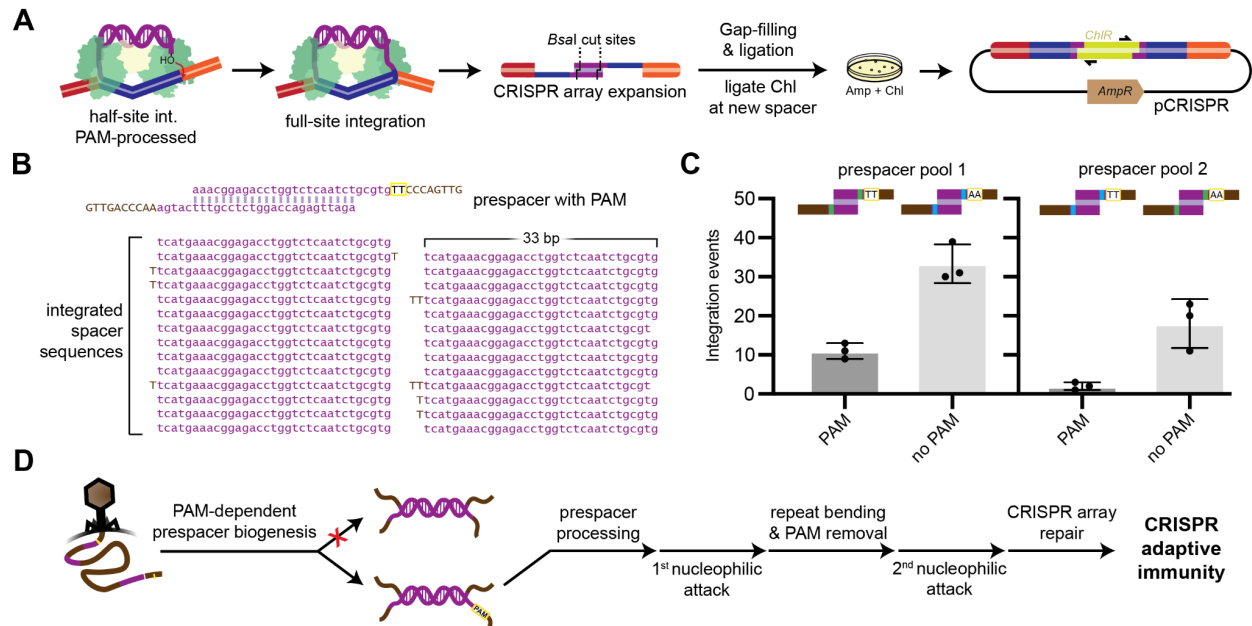


Figure 2-4. In vitro reconstitution of Cas1:Cas2-DEDDh-mediated full-site integration. A) Schematic of in vitro reconstitution of full-site integration. Amp., ampicillin; chl., chloramphenicol. **B)** Fully integrated spacer sequences from a prespacer containing the PAM. Lowercase bases match the prespacer sequence. Uppercase bases are additional, untrimmed nucleotides. **C)** The number of integration events arising from an equimolar pool of prespacers with or without the PAM. Prespacers are distinguished by internal barcodes, which were swapped to remove sequence bias. Data are mean $\pm$ s.d. of three independent biological replicates ($n = 95$ colonies). **D)** Proposed general timeline of CRISPR adaptation.

Although it was hypothesized that delayed PAM trimming aids the complex in orienting the prespacer for integration^{73,75}, no orientation bias was observed *in vitro* (**Supplementary Fig. 2-8C**). While Cas1:Cas2-DEDDh alone is able to distinguish between the PAM and non-PAM sides of the prespacer (**Figs. 2-1D; 2-2C**), it appears

that the complex alone cannot discern the leader- and spacer-side of the repeat, consistent with cryo-EM results of the half-site intermediate, which show no sequence specificity for the leader (**Fig. 2-3D**). We suspect the complex requires additional host factors to correctly orient the spacers in vivo. In *E. coli*, integration host factor (IHF) directs the first nucleophilic attack to the leader-side of the repeat through specific contacts with the leader sequence^{98,103}. Superimposition of the half-site structure and a *Megasphaera* IHF ortholog onto a structure of the complete IHF-containing integration holo complex further implicates the participation of a directing host factor (**Supplementary Fig. 3-9**). In vivo, the system may have higher specificity for the leading integration target site and utilize delayed PAM processing as the basis for determining the orientation of integration, as is the case in other CRISPR systems^{73,75,76}.

To assess the effect of the PAM on integration efficiency, an equimolar mixture of PAM-deficient and PAM-containing prespacers, each containing a pair of identifying internal barcodes, was tested for full-site integration. Surprisingly, out of 95 sequenced colonies, we observed significant enrichment (~3-fold) of integration events from the PAM-deficient prespacer (**Fig. 2-4C**). To account for biases resulting from the internal barcode sequences, we generated a second prespacer pool, where barcode pairs were swapped. The second pool also exhibited a significant preference for the PAM-deficient prespacer (**Fig. 2-4C**). Lower integration efficiency from the PAM-containing prespacer in vitro may stem from additional steps required for PAM removal (**Fig. 2-3I**). Moreover, PAM removal is observed following full-site integration, and all spacer sequences are selected according to PAM presence (**Fig. 2-4B**). The reduced apparent efficiency of PAM-containing prespacer insertion in vitro therefore suggests PAM recognition in vivo occurs upstream, during the biogenesis of substrates bound for CRISPR adaptation (**Fig. 2-4D**).

Integration reconstitution experiments with pooled prespacers suggest that the integrase may select substrates prior to prespacer processing, ensuring PAM presence (**Fig. 2-4C**). We were interested in whether the integrase demonstrates similar stringency for another substrate feature, the canonical prespacer duplex. Accordingly, we tested Cas1:Cas2-DEDDh processing after stepwise addition of the PAM-complementary strand. After incubation of Cas1:Cas2-DEDDh with a single-stranded PAM-containing strand, the labeled, PAM-deficient strand was added. In all reactions, even for the prespacer strands with the tolerated 23-nt complementarity region (**Fig. 2-1B**), nonspecific processing of the labeled strand occurs (**Supplementary Fig. 2-8D-F**). Thus, Cas1:Cas2-DEDDh likely performs ruler-guided trimming when the 23-bp prespacer is pre-duplexed and not after delayed addition or search for the complementary strand. The strict requirement for substrate size, strandedness, and PAM-presence has implications for open questions in CRISPR substrate biogenesis.

2.3.5 Selective processing and protection of prespacer sources and models of biogenesis

Integration reconstitution experiments with pooled prespacers suggest the integrase may select substrates before formation of the canonical prespacer. We tested this model by probing Cas1:Cas2-DEDDh processing of potential prespacer precursors, including PAM-containing dsDNA and ssDNA. We found that Cas1:Cas2-DEDDh cannot selectively process blunt-end dsDNA (using either ruler-guided trimming or PAM-dependent trimming) but, surprisingly, can carry out PAM-dependent recognition and 3' overhang trimming on ssDNA substrates, analogous to results with duplex prespacer substrates (**Fig. 2-5A**).

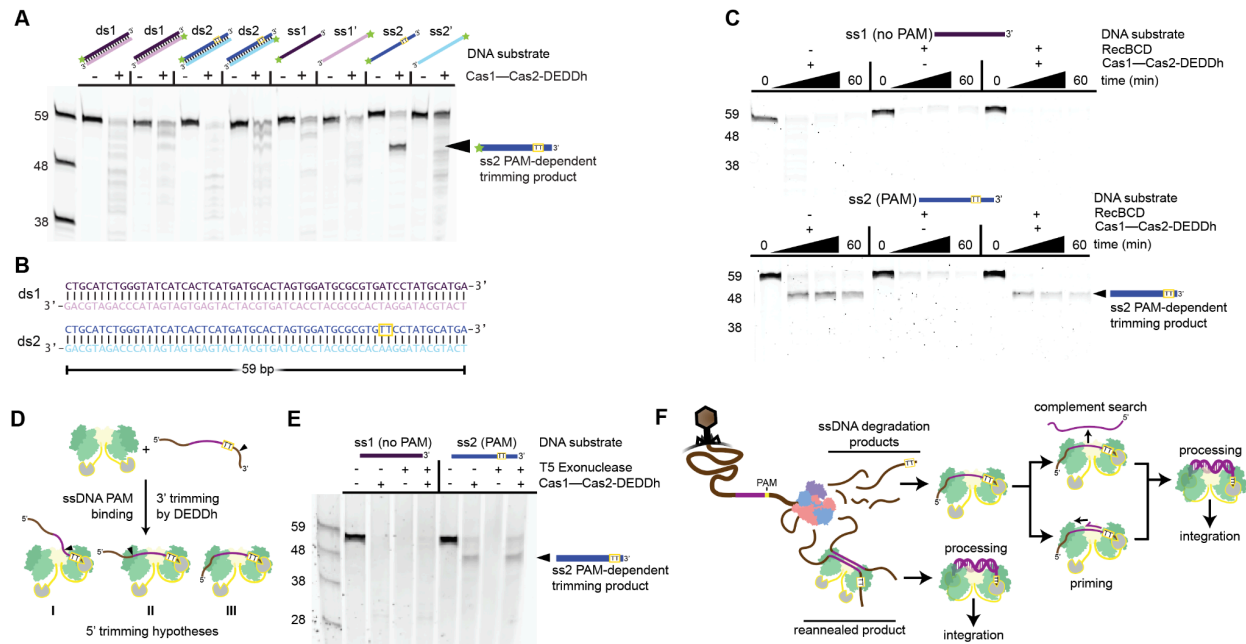


Figure 2-5. Cas1:Cas2-DEDDh selectively performs PAM-dependent trimming on certain ssDNA substrates but not dsDNA. **A)** Processing of PAM-deficient and PAM-containing dsDNA and ssDNA substrates by Cas1:Cas2-DEDDh. **B)** Substrates used in (a), (c), and (d). **C)** Processing of ssDNA substrates in the presence of bipolar helicase RecBCD. **D)** Hypothetical trimming patterns of ssDNA exposed to a 5' exonuclease after capture by the integrase. **E)** Processing of ssDNA with addition of T5 5' exonuclease. **F)** Model of upstream processes in CRISPR adaptation.

If PAM-containing ssDNA substrates are prespacer precursors, the question remains: When is the canonical 23-bp duplex of the prespacer formed during prespacer biogenesis? One plausible hypothesis is that Cas1:Cas2 first binds to ssDNA prespacer precursors and then finds and facilitates pairing with the complementary partner⁷⁵. We tested this hypothesis by assaying Cas1:Cas2-DEDDh processing after stepwise addition of the PAM-complementary strand. After incubation of Cas1:Cas2-DEDDh with a single-stranded PAM-containing strand, the labeled, PAM-deficient strand was added. In all reactions, even for the prespacer strands with

the tolerated 23-nt complementary region, nonspecific processing of the labeled strand occurs (**Supplementary Fig. 2-12**). Thus, Cas1:Cas2-DEDDh likely performs ruler-guided trimming when the 23-bp prespacer is pre-duplexed and not after delayed addition or search for the complement. The strict requirement for substrate size and strandedness has implications for open questions in CRISPR substrate biogenesis.

The preferential biogenesis of spacers from invader DNA is poorly understood. The double strand break resection/ homologous recombination complex RecBCD was implicated in prespacer production¹⁰⁴, but no detailed model exists to explain the formation of CRISPR array-bound duplex prespacers from the asymmetric ssDNA products of RecBCD digestion¹⁰⁵. We were interested in the ability of RecBCD to degrade extended ssDNA bound by the integrase (**Fig. 2-5C**). *E. coli* RecBCD and Cas1:Cas2-DEDDh robustly degrade PAM-deficient substrates in isolation and in combination (**Fig. 2-5C**, top). In PAM-containing substrates, however, Cas1:Cas2-DEDDh completely protects ssDNA from RecBCD-mediated degradation (**Fig. 2-5C**, bottom), despite the increased size of the substrate (59 nt) compared to canonical prespacer substrates (28 nt ssDNA, 33 nt total linear length). RecBCD is a bipolar helicase/endonuclease capable of cleavage of a variety of ds and ssDNA substrates¹⁰⁶. To address the fate of the 5' end of PAM-containing ssDNA, we assayed processing in the presence of T5 exonuclease, a unidirectional 5' ssDNase. DEDDh trims the 3' end in a defined manner, but 5' trimming mechanisms remain unclear. Reasonable hypotheses concerning 5' ssDNA protection include: I) captured ssDNA is trimmed until hindered by the pocket formed by Cas1a' and Cas1b', II) 5' trimming occurs at some intermediate distance, perhaps defined by canonical duplex region, or III) the entire ssDNA is preserved as in **Fig. 2-5C** (**Fig. 2-5D**). Degradation of ssDNA is complete absent the PAM but, surprisingly, Cas1:Cas2-DEDDh completely preserves PAM-containing ssDNA (**Fig. 2-5E**). Protection of the 5' end during exposure to T5 exonuclease is consistent with hypothesis III (**Fig. 2-5D,E**).

2.4 Discussion

Efficient CRISPR adaptive immunity requires coordination between the CRISPR integrase and host nucleases^{75,76}. In this study, we describe mechanisms of prespacer processing and integration in a naturally occurring Cas1:Cas2-DEDDh complex. The trimmer-integrase uses an alternative PAM-processing mechanism compared to the well-studied Cas4 endonuclease^{73,94}. Prior to this work, it was unclear how systems lacking Cas4 process and integrate substrates. Our data suggest that one evolutionary solution to the problem of selecting, protecting, and then removing the PAM is to use Cas1 rather than an accessory protein for initial PAM protection. Sequestration of defined prespacer sizes via substrate gatekeeping ensures that the PAM is present and that its cognate spacer is functional (**Fig. 2-1D**). Once the PAM-containing prespacer is anchored to the host CRISPR array, the PAM is released by the Cas1 'gate' and is

promptly removed. We provide a mechanism explaining which structural cues lead to PAM uncapping and removal (**Fig. 2-3I**). Binding and bending of leader-distal repeat DNA may lead to disengagement of the C-terminal cap which covers and protects nucleotides. DEDDh completely digests the released PAM, generating substrates of the correct size and positioned for second nucleophilic attack. This sequence of events ensures the PAM side integrates second. Bending and unwinding may additionally aid in melting of the repeat strand, which is required for resolution of the post-synaptic complex and concomitant repeat duplication¹⁰⁷. While the high effective local concentration of exonuclease with respect to the bound prespacer conferred by the Cas2-DEDDh fusion in this system likely improves efficiency, we imagine that host exonucleases can function similarly in trans. Cas1:Cas2-DEDDh serves as a general model which describes the role of accessory exonucleases, including those which are not fused to the nuclease, in the diverse CRISPR systems lacking Cas4 (see **Supplementary Fig. 2-10** for comparisons to *E. coli* Cas1:Cas2).

Recent studies indicated accessory proteins can coordinate with Cas1:Cas2 to process prespacers. DnaQ and ExoT have previously been shown to process PAM-containing substrates asymmetrically when provided in concert with Cas1:Cas2 *in vitro*^{75,76}, establishing the directionality of integration. Recent *in vivo* work demonstrated that other accessory exonucleases can substitute for DnaQ and ExoT activities to carry out prespacer processing⁹⁵. These findings suggest that Cas1:Cas2 can flexibly coordinate with various accessory proteins. As a general model, Cas1:Cas2-DEDDh provides insight into the elegant mechanism by which nonspecific processing enzymes and Cas1:Cas2 preserve the *self* vs. *non-self* distinction. These findings advance our understanding of how prespacers are processed and selected for spacer acquisition. We anticipate that our results will be applicable to CRISPR-based technologies which seek to repurpose Cas1:Cas2 for molecular recording and information storage, applications challenged both by reliance on host factors such as exonucleases and by uncertainty in prespacer selection¹⁰⁸⁻¹¹⁰.

While this report represents an advance in our understanding of downstream steps, the upstream biogenesis of CRISPR substrates remains mysterious. Unexpectedly, experimental data suggest lower integration efficiency from PAM-containing prespacers and a preference for preformation of the canonical duplex³⁵. We hypothesize that PAM selection may occur early during prespacer biogenesis, prior to the formation of canonical prespacer substrates. Pool experiments prompted exploration of the capture and selective processing of prespacer precursors, revealing PAM-dependent trimming of ssDNA substrates. Furthermore, only PAM-containing ssDNA are protected from RecBCD and T5 exonuclease degradation, supporting a potential prespacer biogenesis pathway from PAM-containing ssDNA precursors^{111,112}. Interestingly, Cas1:Cas2-DEDDh requires a pre-formed duplex of a defined size for ruler-guided processing to generate competent substrates for integration. Stepwise addition of the two prespacer strands with the correct complementary region

precluded ruler-guided processing into competent integration substrates³⁵. These results weaken the ‘complement search’ model for prespacer biogenesis, which suggests single stranded DNA derived from foreign sources are captured independently by the integrase complex. Alternatively, Cas1:Cas2 may recognize PAM-containing prespacer-like motifs as DNA reanneals behind repair complexes implicated in CRISPR adaptation such as RecBCD and AddAB, as was recently suggested³³. We note that RecBCD- or AddAB-mediated degradation is asymmetric, meaning ssDNA products exit the complex at differential rates, disadvantaging Cas1:Cas2 as complements diffuse away from integrase-PAM ssDNA complex. We advance two alternative models to guide future efforts to determine the molecular mechanism of CRISPR integration substrate biogenesis (**Fig. 2-5F**). First, the complement may be synthesized after PAM-dependent ssDNA capture. CRISPR-associated primase/polymerases (CAPPs) were shown to associate with the integrase,¹¹³ though the complete functional scope of CAPPs or analogous host factors during adaptation remains unclear. Second, Cas1:Cas2 may coordinate with RecBCD or AddAB to organize prespacer capture, perhaps via direct contacts between Cas1 and the terminal end of the recombination complex.¹¹³ We note that during infection, RecBCD/AddAB ssDNA products rapidly accumulate, titrating single-stranded DNA binding proteins (SSB). In limiting unbound SSB environments, nicked degradation products were shown to reanneal behind the RecBCD complex^{114,115}, presenting a potential motif for prespacer recognition (**Fig. 2-5F**, bottom). Recently, the 5’ ssDNA exonuclease RecJ and helicase but not endonuclease activity from the RecBCD complex were shown to enhance prespacer biogenesis in primed adaptation, lending support to this model⁹⁵. Future experimentation might translate these data and insights from this work into a more complete mechanistic understanding and total in vitro reconstitution of naive CRISPR adaptation. The precise mechanistic details of these proposals remain unclear. Future experiments might utilize the compact trimmer-integrase presented here to probe open questions in CRISPR substrate biogenesis and achieve total in vitro reconstitution of naive CRISPR adaptation.

2.5 Methods

2.5.1 Plasmid construction and DNA substrate preparation

To make the target integration plasmid pCRISPR, the leader and the first 3 repeats and spacers of the CRISPR array were ordered as two DNA fragments, which were amplified by PCR and inserted into a pUC19 backbone by Gibson Assembly. DNA oligos used in this study were ordered from Integrated DNA Technologies (IDT). Prespacers and the half-site substrates were formed by heating at 95 °C for 5 min and slow cooling to room temperature in HEPES hybridization buffer (20 mM HEPES, pH 7.5, 25 mM KCl, and 10 mM MgCl₂). For the half-site substrate, hybridization was

carried out with a 1.5-fold excess of the two shortest strands and 1.25-fold excess of the second-largest strand and purified on a 8% native PAGE gel.

2.5.2 Cloning, expression, and purification

The *Megasphaera* NM10_related Cas1 and Cas2-DEDDh genes were codon optimized for *E. coli* expression, ordered as G-blocks, PCR amplified, and cloned separately into a pET-based expression vector with an N-terminal 10xHis-MBP-TEV tag. After transformation into chemically competent Rosetta cells, cells were grown to an OD₆₀₀ of ~0.6 and induced overnight at 16 °C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), and 10% glycerol). After lysis by sonication and clarification of lysate by centrifugation, the supernatant was incubated with Ni-NTA resin (Qiagen). The resin was washed with wash buffer (20 mM HEPES, pH 7.5, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol) and protein was eluted with wash buffer supplemented with 300 mM imidazole. After overnight digestion with TEV protease, the salt concentration was diluted to 300 mM NaCl using ion-exchange buffer A (20 mM HEPES, pH 7.5, 1 mM TCEP, and 5% glycerol) and run through a tandem MBPTrap column (GE Healthcare) and HiTrap heparin HP column (GE Healthcare) to remove the MBP and bind the protein onto the heparin column. The protein was eluted with a gradient from 300 mM to 1 M KCl, concentrated, and purified on a Superdex 200 (16/60) column with storage buffer (20 mM HEPES, pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol). The same purification protocol was used for Cas1 and Cas2-DEDDh (WT and D132A mutant).

2.5.3 Processing assays

Processing assays were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM MgCl₂, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 μM) and Cas2-DEDDh (2 μM) were pre-complexed for 30 min. at 4°C before addition of fluorescent DNA substrate (312.5 nM) and reacting for 2 hours at 37°C. Reaction was quenched by addition of 2 vol. quench buffer (95% formamide, 30 mM EDTA, 0.2% SDS, and 400 μg/mL heparin) and heating at 95 °C for 4 min., before analysis on a 14% urea-PAGE gel. Reactions were visualized by Typhoon FLA gel imaging scanner and quantification of intensities was performed with ImageQuantTL 8.2. The percent processing activity is quantified as the ratio of the final product bands' intensity to the total intensity of all bands in the lane.

2.5.4 Cryo-EM data acquisition

Cas1-Cas2-DEDDh DNA complexes were formed by mixing 50 μM Cas1, 50 μM Cas2-DEDDh, and 12.5 μM prespacer or half-site DNA and dialyzing for 2 h using a

Slide-A-Lyzer MINI Dialysis Device at room temperature. The complex was concentrated to varying concentrations of Cas1:Cas2-DEDDh (see **Supplementary Table 2-1**) and purified over a Superose 6 Increase 10/300 GL column. Samples were frozen using FEI Vitrobot Mark IV, cooled to 8°C at 100% humidity. Depending on the sample (see **Supplementary Table 2-1**), either carbon 2/2 300 mesh C-flat grids (Electron Microscopy Sciences CF-223C-50) or 1.2/1.3 300 mesh UltrAuFoil gold grids (Electron Microscopy Sciences #Q350AR13A) were glow discharged at 15 mA for 25 s using PELCO easyGLOW. In all cases, a total volume of 4 µL sample was applied to the grid and immediately blotted for 5 s with a blot force of 8 units. Micrographs were collected on a Talos Arctica operated at 200 kV and x36,000 magnification (1.115 Å pixel size), in the super-resolution setting of K3 Direct Electron Detector. Cryo-EM data was collected using SerialEM v.3.8.7 software. Images were obtained in a series of exposures generated by the microscope stage and beam shifts.

2.5.5 Cryo-EM data processing

All datasets were collected with varied tilt angle, amount of movies, and defocus range (see **Supplementary Table 2-1** and **Supplementary Figs. 2-3, 2-4, 2-5**). Data processing was performed in cryoSPARC v3.2.0, v3.3.1, and v4.1.1¹¹⁶. Movies were corrected for beam-induced motion using patch motion correction, and contrast transfer function (CTF) parameters were calculated using patch CTF.

PAM-deficient prespacer bound Cas1:Cas2-DEDDh map was obtained by an iterative process. In the first round 569 particles were picked manually from 37 micrographs and submitted for Topaz training¹¹⁷. The resulting Topaz model was used to pick particles from the micrographs, and a total of 460,631 particles were extracted with bin factor 2, and applied to 2D classification. Following the selection of the best classes, 410,757 particles were used for ab initio reconstruction and subsequent heterogeneous refinement, with three classes. All particles were used for non-uniform (NU) map refinement¹¹⁸, and an initial complex map was obtained. Following 2D classification of particles from the initial NU refinement model 38,342 particles from the classes with isotropic orientations were selected and subject to the second round of Topaz training. A new Topaz model was used with a total of 956 curated micrographs, and the entire process was repeated twice with particles from the best heterogeneous refinement class for subsequent NU refinement and Topaz training. The final map with the best electron density for the PAM-deficient prespacer bound Cas1:Cas2-DEDDh complex was obtained from 461,266 particles and was refined with non-uniform refinement to 3.1 Å.

For the PAM-containing prespacer bound Cas1:Cas2-DEDDh, a single round of Topaz training was applied. Following the initial exposures curation which yielded 591 best quality micrographs, 6,302 particles were manually picked and subject to Topaz training job. The Topaz model was applied to an expanded set of 1,184 curated micrographs, and resulted in extraction of 3,101,776 particles. After ab initio

reconstruction and heterogenous refinement of the particles, with three classes, the 1,420,721 particles set constituting the best class were processed with non-uniform refinement. As a result, a 2.9 Å density for PAM-containing prespacer bound Cas1-Cas2 complex was obtained.

For resolving the DEDDh density in the latter dataset, the ab initio class particles used for the latter density reconstruction, 1,331,357 in total, were applied to a 2D classification job, and 228,220 particles were selected in classes with apparent DEDDh density. Following ab initio refinement with three classes, particles from the best class were subject to another round of 2D classification, and 109,912 particles with more pronounced DEDDh density were selected, and re-extraction with 320 pixel box size (in all other cases, 480 pixel boxes were used for the extraction jobs). As a result of the final 2D classification round, 49,560 particles with the best DEDDh density were selected, re-extracted with standard box dimensions, and subject to ab initio refinement, with one class, and non-uniform refinement. As a result, a 3.5 Å complex map with the DEDDh exonuclease density was obtained, with a total of 49,383 particles used for reconstruction.

For half-site DNA-bound Cas1:Cas2-DEDDh, the Topaz model from the PAM-containing prespacer was applied to 2,810 micrographs selected after manual curation. The 2,448,888 resultant particles were subdivided using 2D classification, and the 25 best classes were selected, resulting in 1,836,610 particles. These particles were subjected to ab initio reconstruction with three classes. The best class containing 1,048,353 particles was refined using non-uniform refinement to yield to the 3.1 Å half-site map.

To observe DNA dynamics in the Cas1:Cas2-DEDDh half-integration complex, we performed 3-D variability analysis (3DVA)¹⁰⁰ on a subset of particles selected and refined from 2-D classification with DNA visible on the leader-distal side of the complex (1,048,353 particles). The filter resolution was 6 Å and the number of modes was 3. Next, the 3DVA output mode was set to cluster and the number of clusters to 20. Each resulting cluster was individually inspected and two clusters representing maxima of DNA motion along the pitch axis were chosen. The linear structure was derived from 32,722 particles and was subjected to non-uniform refinement to give the final 4.1 Å map. The bent structure resulting from initial 3DVA clustering was improved by repetition of the 3DVA workflow with the complete particle set obtained by Topaz picking, then selection and non-uniform refinement of the cluster representing leader-distal DNA in the most bent conformation (53,545 particles total), yielding the final 3.9 Å map.

2.5.6 Model building and refinement

The initial models of the Cas1 and Cas2-DEDDh were obtained with the AlphaFold 2 program⁵⁰. To build the model of Cas1:Cas2-DEDDh bound to a prespacer with TT PAM complex, the predicted Cas1, Cas2 monomers were docked independently into the corresponding map with the fitmap tool in UCSF ChimeraX v1.2.5¹¹⁹. The DNA models were built de novo. The complex model was refined using rounds of real-space refinement and rigid body fit tools in Coot v0.9.4.1¹²⁰, and real_space_refine tool in Phenix v1.19.2-4158¹²¹, using secondary structure, Ramachandran, and rotamer restraints. This complex model served as an initial model for other Cas1:Cas2 structures, which were refined in an analogous manner.

2.5.7 Ligation assays with pCRISPR integration target plasmid

Ligation assays were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM MgCl₂, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 μM) and Cas2-DEDDh (2 μM) were pre-complexed for 30 min. at 4°C before addition of DNA substrate (312.5 nM) and integration target pCRISPR (20 ng/mL, ~10 nM) and reacting for 2 hours at 37°C. Reaction was quenched with 0.4% SDS and 25 mM EDTA, treated with proteinase K for 15 min at room temperature, and then treated with 3.4% SDS. Reactions were analyzed on a 1.5% agarose gel and visualized by Typhoon FLA gel imaging scanner.

2.5.8 Full-site integration assays

Integration assays (50 μL reactions) were conducted in integration buffer (20 mM HEPES, pH 7.5, 125 mM KCl, 10 mM MgCl₂, 1 mM DTT, 0.01% Nonidet P-40, and 10% DMSO). Cas1 (4 μM) and Cas2-DEDDh (2 μM) were pre-complexed for 30 min. at 4°C before addition of DNA substrate containing BsaI cut sites (312.5 nM) and reacting for 15 minutes, followed by the addition of the integration target pCRISPR (20 ng/mL, ~10 nM) and incubating for 2 hours at 37°C. Products were purified with a DNA Clean and Concentrator 5 kit (Zymo Research) and eluted with 6 μL water. A gap-filling reaction (20 μL total, 37°C for 30 min) was conducted with the purified integration products as described by a previous study¹⁰¹: 6 μL purified acquisition reaction, 6.5 μL water, 2 μL 10x Taq DNA Ligase buffer (NEB), 2 μL dNTP Solution Mix (10 mM stock, NEB), 2 μL Taq DNA ligase (80 U, NEB), and 1 μL T4 DNA Polymerase (1 U, NEB). Gap-filling reactions purified using the Zymo Research kit and eluted with 6 μL water. A Golden Gate-compatible chloramphenicol selection cassette was generated by PCR with primers encoding BsaI cut sites and purified using the Qiagen MinElute PCR Purification Kit. A Golden Gate cloning reaction was performed with the purified, gap-filled integration products and chloramphenicol selection cassette following a standard BsaI assembly protocol. The products were purified using the Zymo Research kit and eluted with 6 μL water and 1 μL was electroporated into DH10B cells (NEB). Electroporated cells were recovered in 975 μL of LB and plated on LB agar containing

carbenicillin (100 µg/mL) and chloramphenicol (25 µg/mL). Of the surviving colonies, 95 were sequenced using Sanger Sequencing and sequences were analyzed using SnapGene Version 5.0.8.

2.5.9 CRISPR locus bioinformatic analysis

Cas2-DEDDh-containing loci from metagenomic data were identified by determining genomes that contained a CRISPR locus via CRISPRDetect, and coding sequences within 5kb of the array were extracted¹²². A DEDDh HMM model was built from BLAST searches against the NCBI nr database that were manually verified¹²³. The coding sequences were searched against the DEDDh model using hmmsearch with $E < 1 \times 10^{-5}$ ¹²³. Matches that also contained credible hits to Cas1 and neighboring other Cas proteins were shortlisted for this work. A preliminary Cas2-DEDDh model was computed using AlphaFold 2 to aid in structure-building⁵⁰.

2.5.10 Statistics and reproducibility

For biochemical experiments, results represent gels of the highest quality. All experiments were generally performed at least in duplicate, though not in the exact same format. Measurements were taken from distinct samples. For full-site experiments, we typically performed assays in biological triplicate to ensure reproducibility. All data points are displayed on the figure panels.

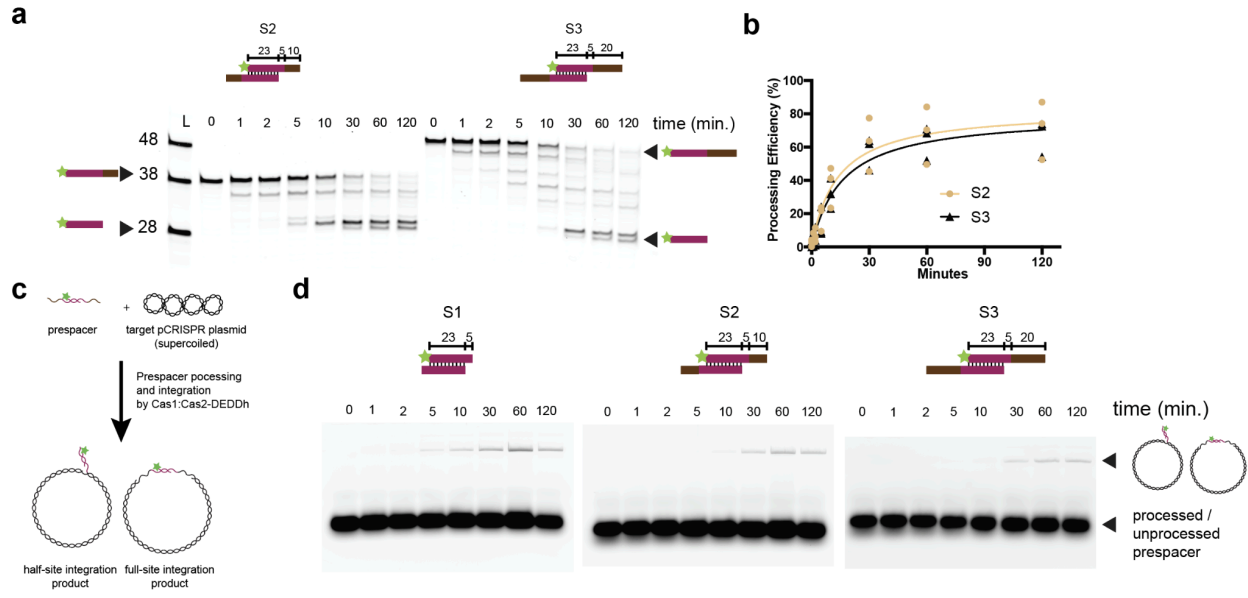
2.5.11 Data availability

Atomic models in the Protein Data Bank (PDB) and the corresponding cryo-EM density maps from the Electron Microscopy Data Bank (EMDB) are available with the following accession codes: Cryo-EM structure of Cas1:Cas2-DEDDh: PAM-deficient prespacer complex PDB-8FY9, EMDB-29561; Cryo-EM structure of Cas1:Cas2-DEDDh: PAM-containing prespacer complex: PDB-8FYA, EMDB-29562; Cryo-EM structure of Cas1:Cas2-DEDDh half-site integration complex: PDB-8FYB, EMDB-29563. Cryo-EM structure of Cas1:Cas2-DEDDh:half-site integration complex linear CRISPR repeat conformation: PDB-8FYC, EMDB-29564; Cryo-EM structure of Cas1:Cas2-DEDDh:half-site integration complex with CRISPR repeat bent conformation: PDB-8FYD, EMDB-29565.

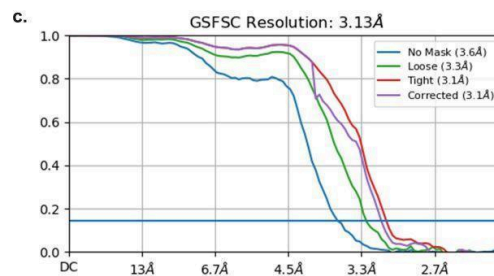
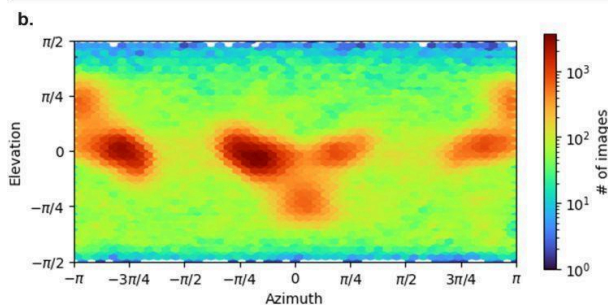
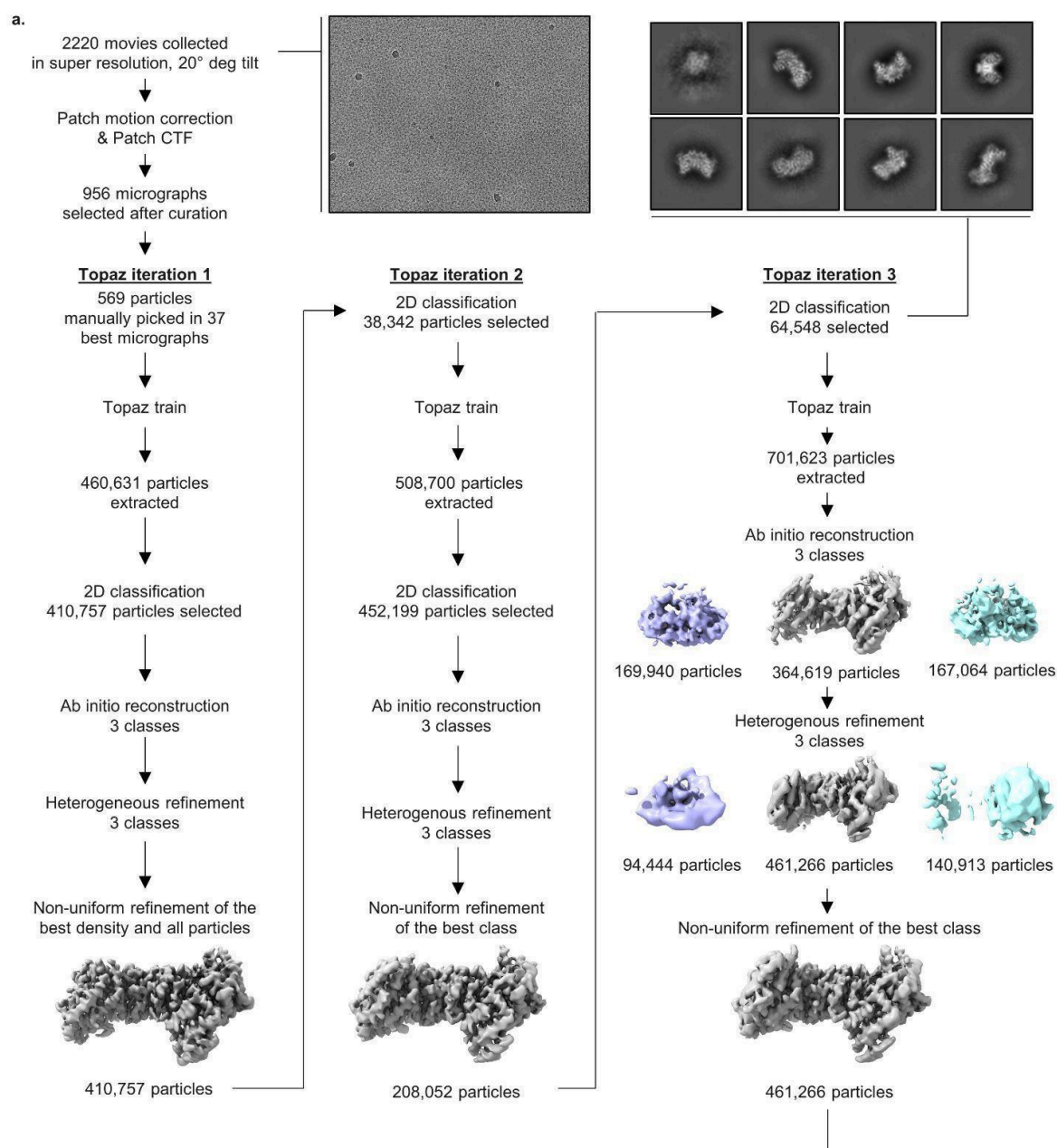
2.6 Supplementary Information

Supplementary Table 1. Cryo-EM Map and Model Validation

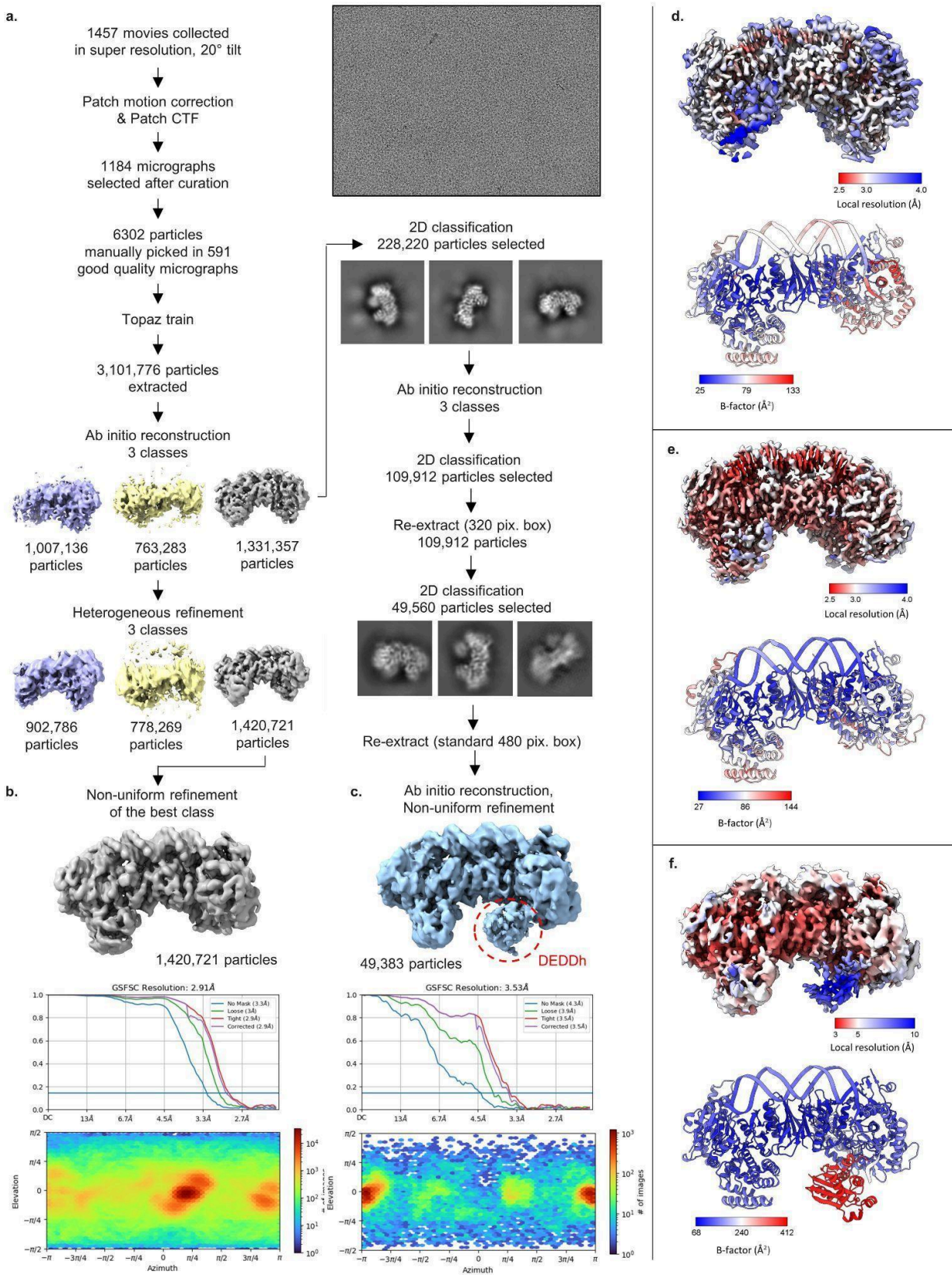
Model	Model 1 (no PAM)	Model 2 (PAM)	Model 3 (PAM-DEDDh)	Model 4 (Half-site)	Model 5 (Half-site linear)	Model 6 (Half-site bent)
PDB accession code	8FY9	8FYA	n/a	8FYB	8FYC	8FYD
EMDB accession code	29561	29562	n/a	29563	29564	29565
Data collection and processing						
Electron Microscope	Talos Arctica					
Electron Detector	K3 Direct Electron Detector					
Magnification	x36,000					
Voltage (kV)	200					
Electron dose (e-/Å²)	50					
Pixel size (Å)	1.115					
Defocus range	0.8 - 2.2 µm	0.0 - 2.2 µm		0.8 - 2.2 µm		
Tilt angle	0°	20°		20°		
Grids	carbon 2/2 300 mesh C-flat grids	1.2/1.3 300 mesh UltrAuFoil gold grids		1.2/1.3 300 mesh UltrAuFoil gold grids		
Complex molarity	8 µM	2.4 µM		2.5 µM		
3D reconstruction						
Raw images	2220	1457	1457	6066	6066	6066
Initial particles	701,623	3,101,776	228,220	1,836,610	1,048,353	1,836,610
Final particles	461,266	1,420,721	49,383	1,048,353	58,475	53,545
Map resolution (Å)	3.13	2.91	3.53	3.14	4.06	3.88
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Map sharpening B-factor (Å ⁻²)	140	138	112.1	165.7	118.7	142.6
Model refinement						
Initial model used	Model 2	Ab initio AlphaFold2 model	Model 2	Model 2	Model 2	Model 2
Model resolution	3.3	3.1	4.1	3.3	4.4	4.3
FSC threshold	0.5	0.5	0.5	0.5	0.5	0.5
Model composition						
Nonhydrogen atoms	10742	11433	12724	13486	14567	13612
Protein residues	1228	1304	1461	1359	1475	1325
Nucleotide	56	59	61	136	145	156
Ligands	0	0	0	0	0	0
B factors (Å²)						
Protein	68.66	55.33	154.31	77.25	179.33	119.59
Nucleotide	76.85	48.17	142.08	167.66	213.59	209.82
R.m.s. deviations						
Bond length (Å)	0.003	0.003	0.003	0.004	0.004	0.003
Bond angles (°)	0.507	0.511	0.647	0.582	0.683	0.582
Validation						
MolProbity score	1.77	1.73	1.85	1.83	1.99	1.85
Clash score	10.84	12.06	15.87	12.15	19.51	13.04
Rotamer outlier (%)	0	0.09	0.08	0	0.25	0
Ramachandran statistics						
Favored	96.62	97.28	97.22	96.51	96.71	96.57
Allowed	3.38	2.72	2.78	3.49	3.22	3.43
Outlier	0	0	0	0	0.07	0
Rama-Z score, whole (r.m.s. Rama-Z)	0.99 (0.24)	1.24 (0.24)	0.14 (0.22)	0.66 (0.23)	-0.60 (0.21)	-0.04 (0.23)
Map CC (box)	0.7	0.68	0.69	0.71	0.82	0.81
Map CC (mask)	0.77	0.78	0.72	0.78	0.74	0.79



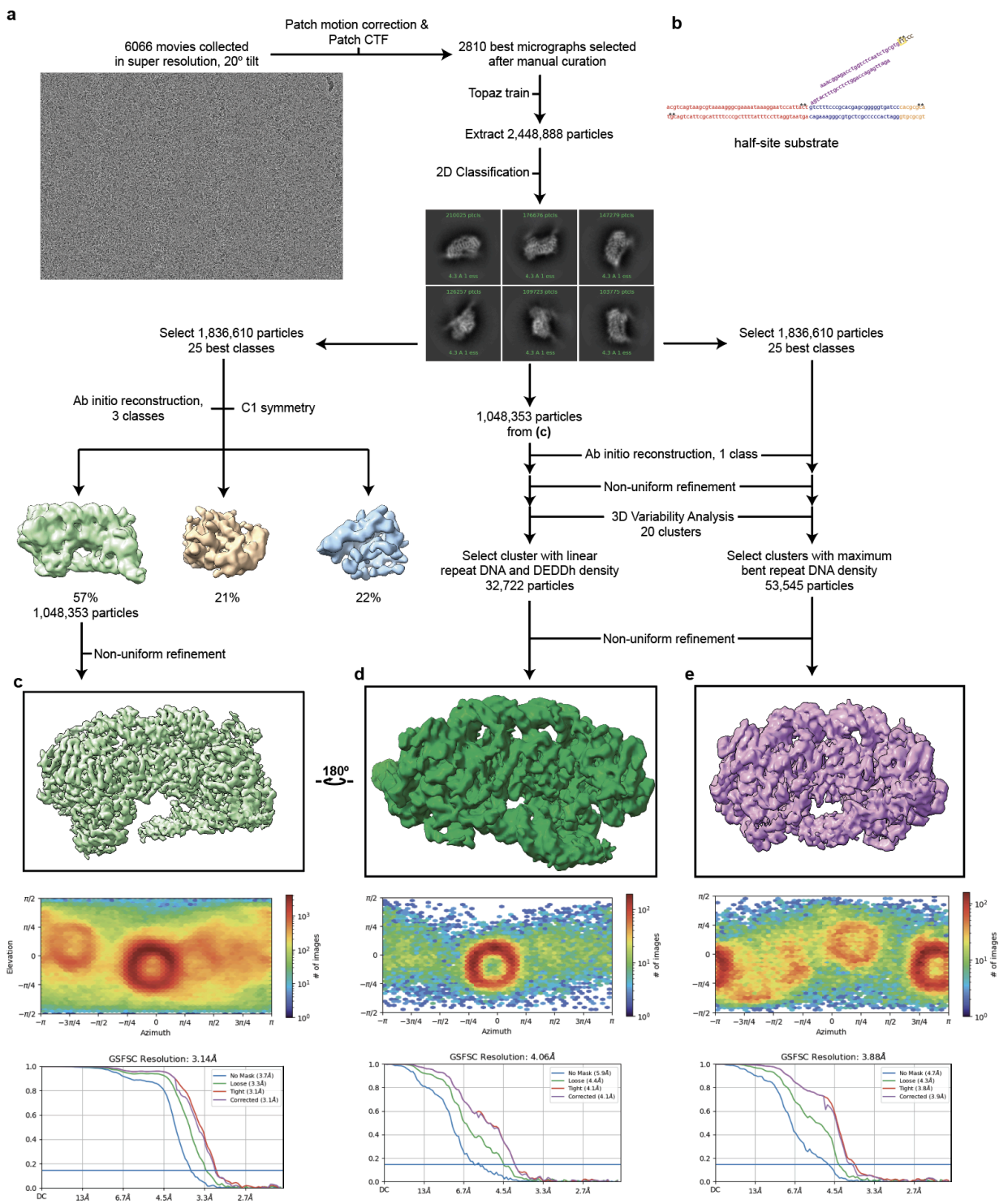
Supplementary Figure 2-2. Kinetics of exonuclease trimming and ligation by Cas1:Cas2-DEDDh *in vitro*. **A)** Time-course reactions of ruler-guided trimming by Cas1:Cas2-DEDDh using substrates S2 and S3 over two hours. **B)** Quantification of time-course reactions of ruler-guided trimming shown in a. Processing efficiency is calculated as the percentage of fully processed product formation at 28-29 nt ($n=3$ biologically independent experiments). **C)** Schematic of *in vitro* ligation assay with integration target pCRISPR plasmid and prespacer substrate. Star indicates 6-carboxyfluorescein (6-FAM) label. **D)** Time-course reactions of ligation assay with substrates from (b) over 2 hours. Note formation of a 6-FAM-labeled pCRISPR plasmid ligation product band.



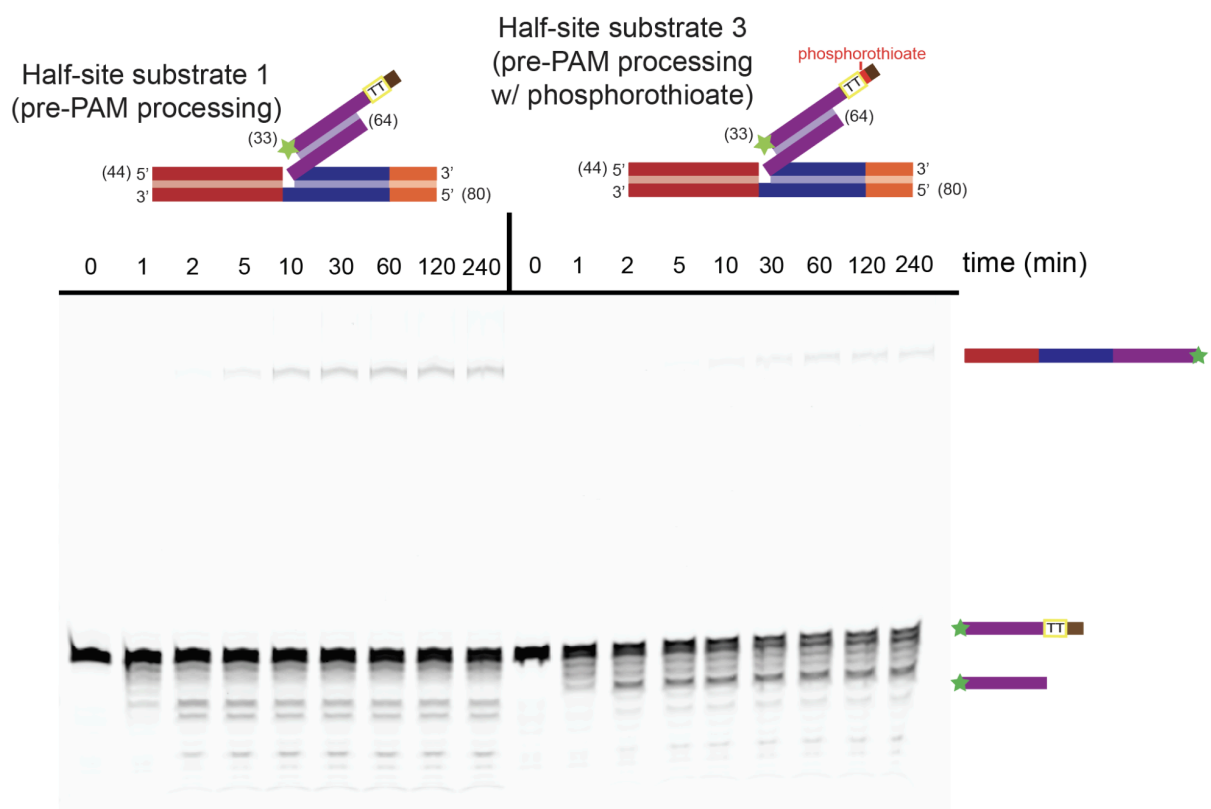
Supplementary Figure 2-3. Flow-chart of the cryo-EM single particle reconstruction of the PAM-deficient prespacer bound Cas1-Cas2/DEDDh. A) Map generation pipeline in cryoSPARC consisting of three Topaz training iterations, including representative 2D class averages and 3D maps resulting from *ab initio* reconstruction and further heterogeneous, non-uniform refinement. **B)** Orientation distribution of the final set of refined particles. **C)** Gold standard FSC curve of the atomic model refined to the final cryoSPARC sharp map.



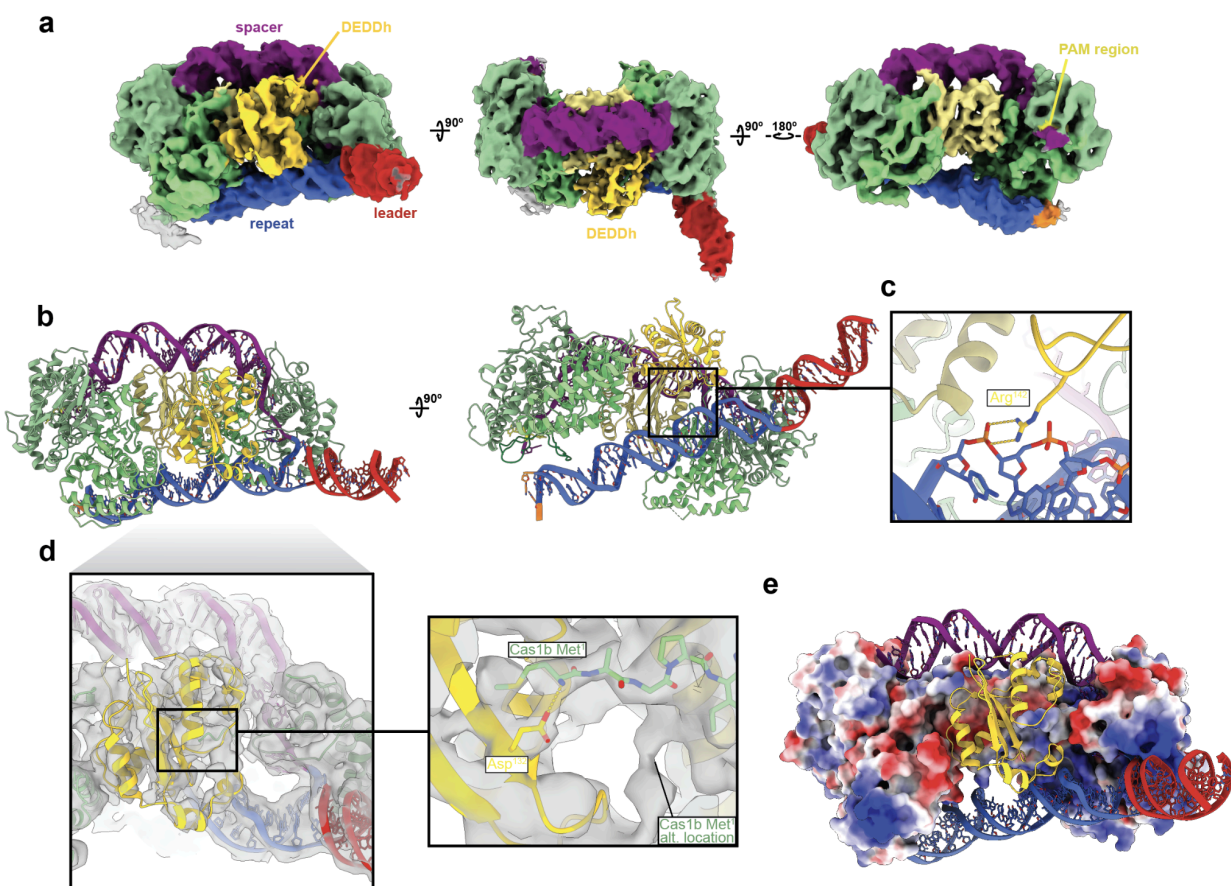
Supplementary Figure 2-4. Flow-chart of the cryo-EM single particle reconstruction of the PAM-containing prespacer bound Cas1-Cas2 and resolution of the DEDDh density. **A)** Cryo-EM data collection parameters and map generation pipeline in cryoSPARC, including representative 2D class averages and 3D maps resulting from *ab initio* reconstruction and further heterogeneous refinement. **B,C)** 3D maps, orientation distribution, and gold standard FSC curve of the final cryoSPARC map for the PAM-only (b) and DEDDh-containing (c) densities. **D-F)** Final sharp maps of Cas1:Cas2-DEDDh bound to prespacers colored according to local resolution as calculated by cryoSPARC, and the final refined models colored with B-factors as calculated by Phenix, for: (d) PAM-deficient density, (e) PAM-containing DEDDh-absent density, and (f) PAM-containing, DEDDh-containing density.



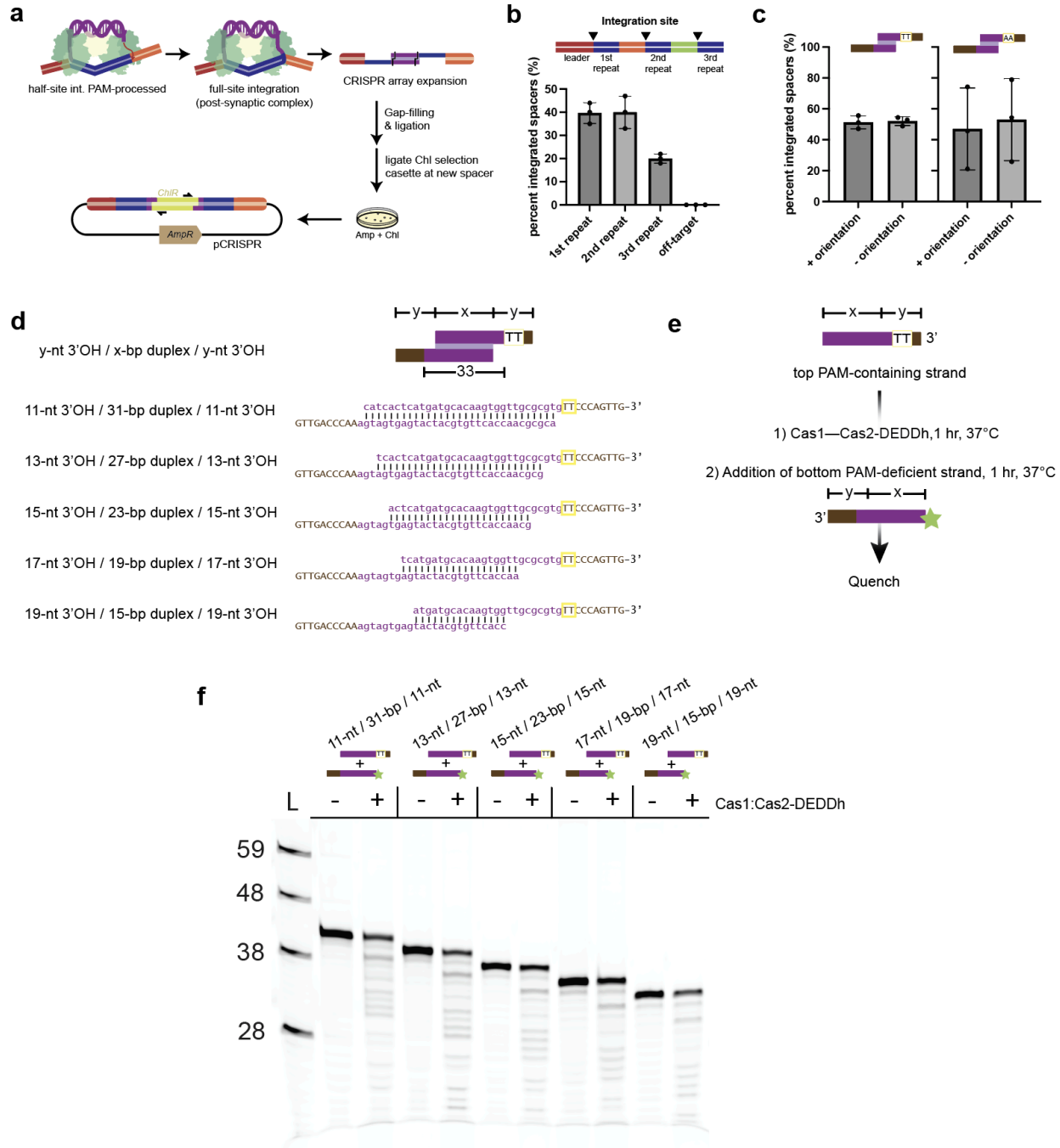
Supplementary Figure 2-5. Flow-chart of the cryo-EM single particle reconstruction of the PAM-containing half-site substrate bound to Cas1-Cas2/DEDDh. A) Cryo-EM data collection parameters and map generation pipeline in cryoSPARC, including representative 2D class averages and 3D maps resulting from *ab initio* reconstruction and further heterogeneous refinement. Different particle stacks were used for generation of the high resolution structure (c), and 3D variability analysis resulting in linear (d) and bent (e) density maps. **B)** Representation of the 4 oligonucleotide half integration construct used in this study. Asterisk indicates a phosphorothioate bond. **C-E)** 3D maps, orientation distribution, and gold standard FSC curve of the atomic model refined to the final cryoSPARC sharp map for each half-site structure.



Supplementary Figure 2-6. Incomplete resistance of phosphorothioate PAM nucleotides to DEDDh. Processing assay (4 hour) with substrates shown in above. Substrates are schematized and TT PAM is indicated, along with phosphorothioation. Processing products are depicted on the right.

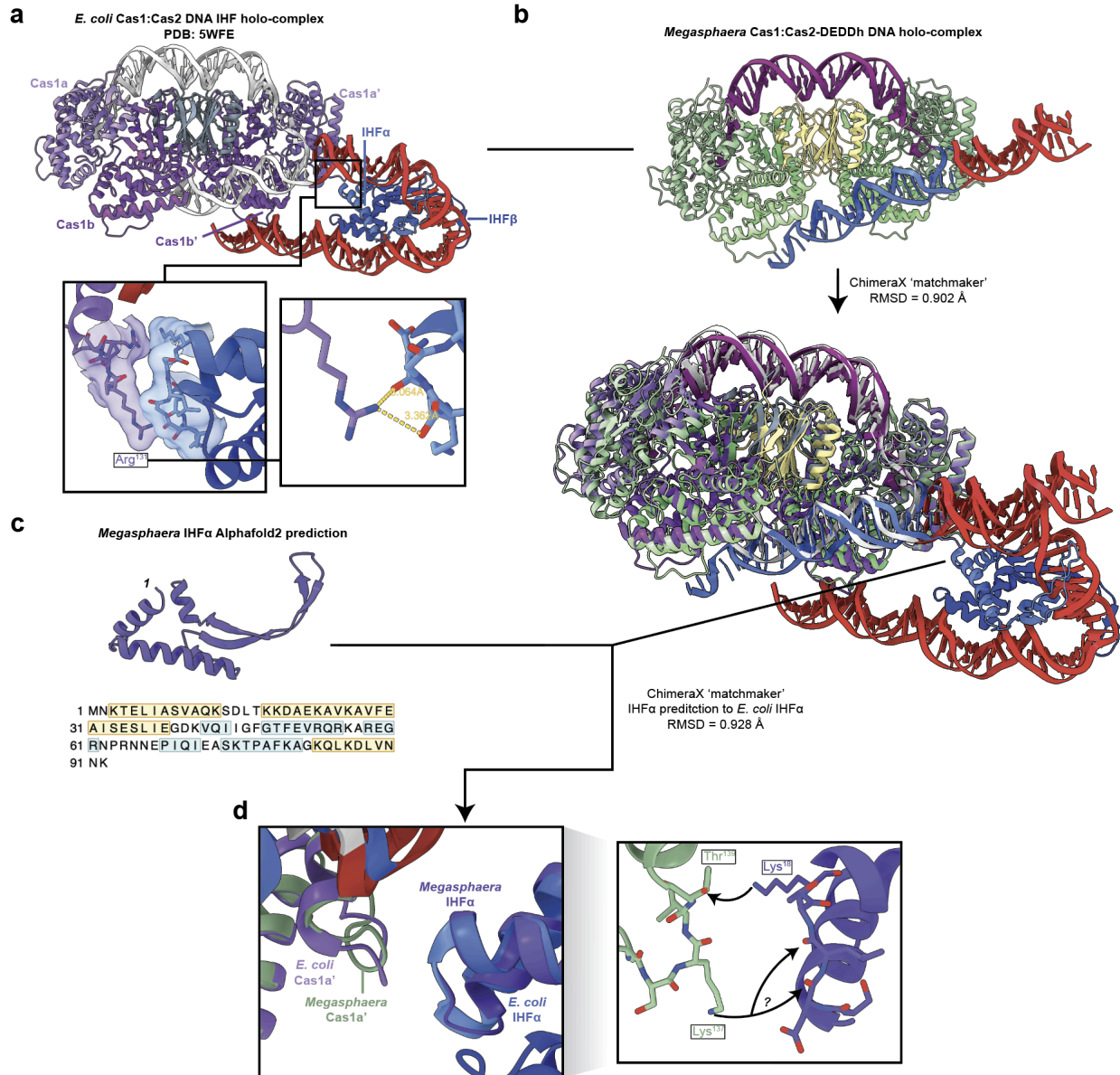


Supplementary Figure 2-7. Molecular detail of Cas1:Cas2-DEDDh in the linear CRISPR repeat DNA conformation. **A)** Orthogonal views of the final sharpened cryo-EM densities for Cas1:Cas2-DEDDh bound to half-site DNA intermediates containing a phosphorothioated TT PAM, colored to demonstrate domain locations. **B)** Non-PAM side and bottom views of Cas1:Cas2-DEDDh bound to half-site intermediate DNA. **C)** Detail depicting potential interaction between Arg¹⁴² of DEDDh and CRISPR repeat DNA phosphate backbone. **D)** Non-PAM side view of the DEDDh domain with the sharp map superimposed (threshold: 0.05). Right, detail in the catalytic pocket of DEDDh, with Asp¹³² shown in close proximity to an extension attributed to the N-terminus of Cas1b. **E)** Cas1:Cas2-DEDDh linear structure, with protein surfaces except the DEDDh domain colored according to electrostatic potential.

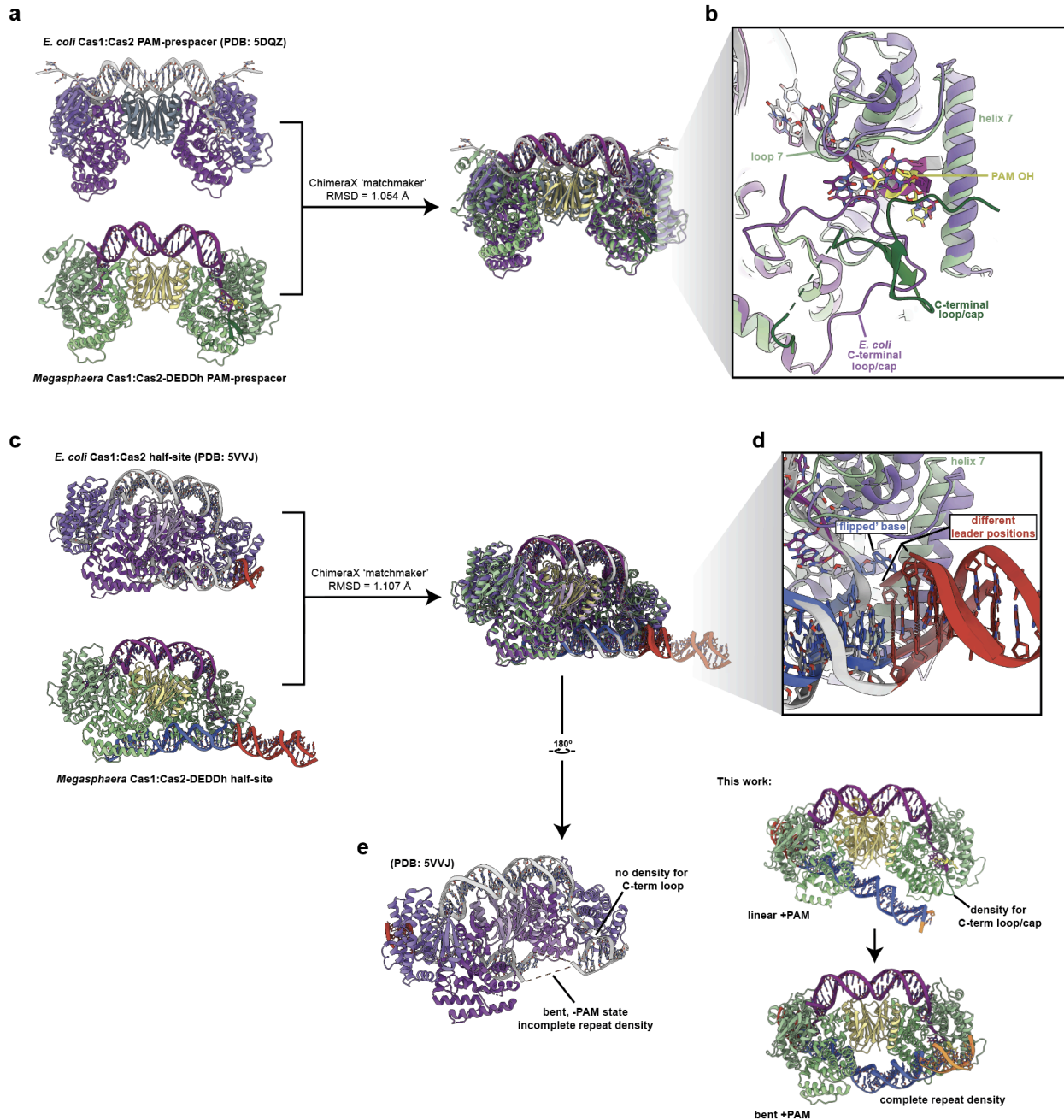


Supplementary Figure 2-8. Integration site preference, orientation bias, and duplex requirements of Cas1—Cas2-DEDDh. **A)** Schematic depicting experimental workflow for full-site integration reconstitution. **B)** Integration events at each integration site from sequenced clones. Integration sites are depicted by arrows at repeat borders. The mean and standard deviation of three independent replicates are shown ($n = 95$ colonies). **C)** Left, orientation of spacer insertion from prespacer containing PAM motif, + orientation orients original TT PAM containing end toward the leader, - orientation orients original TT PAM containing end away from the leader. Right, Orientation of spacer insertion from control prespacer without PAM motif (the TT PAM from left is replaced with AA). The mean and standard deviation of three independent replicates are shown ($n = 95$ colonies). **D)** Prespacer substrates with varying duplex lengths used in processing assay. The duplex and overhang lengths are indicated, and the TT PAM motif is

boxed. **E)** Schematic of processing assay with stepwise addition of the top and bottom prespacer strands of the substrates shown in d. The unlabeled top PAM-containing strand is incubated first with Cas1:Cas2-DEDDh followed by delayed addition of the labeled bottom PAM-deficient strand. **F)** Processing assay with stepwise addition of the top and bottom prespacer strands of the substrates shown in a.



Supplementary Figure 2-9. Comparison of Cas1:Cas2-DEDDh bound to the half-site intermediate with the complete *E. coli* IHF-containing integration complex. A) *E. coli* Cas1:Cas2 DNA IHF holo-complex (PDB: 5WFE). Subunit identities are indicated above. Inset boxes highlight the interface between Cas1a and IHF. **B)** Representation of the pipeline used to compare *Megasphaera* and *E. coli* integrases. The experimental structure solved in this work was superimposed with the *E. coli* complex using ChimeraX matchmaker⁵⁴. **C)** The sequence of a *Megasphaera* IHFα ortholog was found by protein BLAST of *E. coli* IHFα sequence in the *Megasphaera* sp. An286 genome (Taxon ID: 1965622). The AlphaFold 2 sequence of the highest confidence hit is shown along with the protein sequence, with secondary structures indicated. **D)** The *Megasphaera* IHFα was superimposed with the *E. coli* IHFα to approximate the interface in a hypothetical *Megasphaera* complex. Inset shows the structures overlaid (left) and with only *Megasphaera* proteins shown (right). Right, potential interactions at the interface of *Megasphaera* Cas1 and IHFα.



Supplementary Figure 2-10. Comparison of Cas1:Cas2-DEDDh to *E. coli* Cas1:Cas2 structures. A) *E. coli* Cas1:Cas2 bound to a PAM-containing prespacer (PDB: 5DQZ) was compared to *Megasphaera* Cas1:Cas2-DEDDh bound to a PAM-containing prespacer. The experimental structure solved in this work was superimposed with the *E. coli* complex using ChimeraX matchmaker⁵⁴. **B)** Inset showing detail of the superimposition at the PAM binding site. The C-terminal loop in *E. coli* Cas1:Cas2 is oriented differently but covers the PAM in an analogous manner. **C)** *E. coli* Cas1:Cas2 bound to a PAM-absent half-site DNA (PDB: 5VVJ) was compared to *Megasphaera* Cas1:Cas2-DEDDh bound to a PAM-containing half-site DNA. **D)** Inset depicting differences at the integration strand junction. **E)** Comparison of the unintegrated side of both holoenzymes.

Chapter 3

Genome integrity sensing by the broad-spectrum Hachiman antiphage defense complex

This chapter is adapted from¹²⁴: Tuck, O.T.*, Adler, B.A.*, Armbruster, E.G., Lahiri, A., Hu, J.J., Zhou, J., Pogliano, J., and Doudna, J.A. (2024). Genome integrity sensing by the broad-spectrum Hachiman antiphage defense complex. *Cell* 187, 6914–6928.e20.

* these authors contributed equally

O.T.T. and B.A.A. conceived of this project. B.A.A. performed phylogenetic analyses with assistance from O.T.T. B.A.A. and A.L. performed molecular cloning, phage challenge assays, and pharmacological assays. O.T.T. and J.J.H. performed protein purification and biochemical experiments with assistance from J.Z. Cryo-EM data were collected by O.T.T. Cryo-EM data processing was performed by O.T.T. and J.J.H. E.G.A. performed live-cell microscopy experiments. This work was supervised by J.P. and J.A.D. The manuscript was written by O.T.T., B.A.A., and J.A.D. All authors contributed to reviewing and editing this manuscript.

3.1 Abstract

Hachiman is a broad-spectrum antiphage defense system of unknown function. We show here that Hachiman is a heterodimeric nuclease-helicase complex, HamAB. HamA, previously a protein of unknown function, is the effector nuclease. HamB is the sensor helicase. HamB constrains HamA activity during surveillance of intact dsDNA. When the HamAB complex detects DNA damage, HamB helicase activity activates HamA, unleashing nuclease activity. Hachiman activation degrades all DNA in the cell, creating ‘phantom’ cells devoid of both phage and host DNA. We demonstrate Hachiman activation in the absence of phage by treatment with DNA-damaging agents, suggesting that Hachiman responds to aberrant DNA states. Phylogenetic similarities between the Hachiman helicase and enzymes from eukaryotes and archaea suggest deep functional symmetries with other important helicases across domains of life.

3.2 Introduction

Helicases are diverse and ancient motor proteins which comprise well over 1% of all protein-coding genes. Helicases use energy from nucleoside triphosphate (NTP) hydrolysis to manipulate and traverse nucleic acids, regulating the flow of information in the cell^{125,126}. Although best known for their roles in unwinding the DNA double helix during genome replication¹²⁷, helicases perform other essential functions in recombination and repair¹²⁸, transcription¹²⁹, RNA splicing¹³⁰, mRNA decay¹³¹, ribosome biogenesis¹³², and translation initiation¹³³. Helicases are also important components of immune systems. Many immune-associated helicases belong to the superfamily 2 (SF2) class¹³⁴, which contains two bacterial RecA-like domains. In metazoans, SF2 helicases such as Dicer in RNA interference (RNAi) and RIG-I in the type I interferon (IFN) response detect the accumulation of viral double-stranded RNA (dsRNA) in the cytoplasm and unleash immune responses^{135,136}. Dysregulation of helicase function is associated with a range of diseases in humans¹³⁷. In prokaryotes, Type I and VI CRISPR-Cas adaptive immune systems use the Cas3 and DinG SF2 helicases, respectively, to degrade foreign DNA upon CRISPR RNA-guided target recognition^{138,139}. Restriction-modification (RM) systems frequently encode helicases, apparently to facilitate DNA search^{140–142}. RM and Type I CRISPR-Cas systems are the most abundant immune systems in prokaryotes⁵⁸. Across all kingdoms of life, helicases protect the host by sensing pathogen-associated molecular patterns (PAMPs) and mounting an immune response.

Recent work revealed the diversity of immune systems used by prokaryotes against their viruses, termed bacteriophage or phage^{41,43,58}; 152 classes of antiviral defense systems have been described to-date¹⁴³. Many newly discovered antiphage defense systems encode SF2 helicases, including Zorya⁴¹, Shango¹⁴⁴, Azaca¹⁴⁴, DISARM¹⁴⁵, Hna¹⁴⁶, BREX¹⁴⁷, and Dpd¹⁴⁸, among others. In some cases, helicase-containing immune systems embed within other defense architectures to confer new function, as was recently shown for the Hma system¹⁴⁹. Of the helicase-containing defense

systems, the majority have no known mechanism of action. One such system is Hachiman, a two-gene locus containing a short domain of unknown function (DUF1837), HamA, and a SF2 Ski2-like helicase, HamB⁴¹. Although present in >5% of prokaryotic genomes and capable of robust protection against phylogenetically distinct phages^{41,143}, molecular mechanisms governing Hachiman and many related helicase-containing immune systems remain unknown.

Here we show that despite its homology to RNA helicases, HamB is a DNA helicase that activates DNase activity of HamA upon detection of damaged DNA. Cryo-EM structures show how the HamAB complex binds DNA in different modes to facilitate immunity. Helicase ‘ratcheting’ by HamB upon substrate recognition modulates the HamAB interface, leading to HamA activation and indiscriminate degradation of DNA. In situ fluorescence microscopy shows that Hachiman clears both host and phage DNA simultaneously, creating ‘phantom’ cells devoid of genetic material. The observation of Hachiman activation in the absence of bacteriophage but in the presence of DNA-damaging agents suggests that Hachiman responds to DNA damage that accumulates during cell stress. Biochemical and structural data imply ATP-bound HamAB contacts intact DNA, enabling detection of genome integrity and activation of the HamA effector when DNA damage surpasses a normal threshold. HamA nuclease activity may create additional sites for Hachiman binding and activation, leading to amplification of the immune signal and culminating in restriction of phylogenetically diverse phages.

HamB helicase domain organization and its ability to regulate the HamA effector enables controlled activation that may be a principle of other helicase-containing defense pathways. Like Hachiman, other defense systems may act in response to cell stressors including but not limited to phage infection.

3.3 Results

3.3.1 *Hachiman is broad-spectrum helicase defense system*

Helicases are common components of immune systems in eukaryotes^{5,134}. This is also true for prokaryotic immune systems. Of the ~150 prokaryotic defense systems cataloged in DefenseFinder¹⁴³, 18 contain an SF1/SF2 helicase, comprising nearly 20% of non-restriction-modification (RM) defense loci identified in RefSeq (**Fig. 3-1A**)¹⁵⁰. Using 95 defense system helicases and 236 well-characterized representative SF1/SF2 helicases¹³⁸, we performed a phylogenetic analysis of the core helicase domain (**Fig. 3-1B**, see Methods). We assigned 20 of the 25 helicases to an established helicase subfamily, spanning 7 subfamilies. Helicases from Shango (SngC), BREX (BrxHI), Druantia (DruE) and Dpd (DpdJ) formed a potentially distinct clade of SF2 helicase family defined by anti-phage immune functions. A helicase from DISARM (DrmA) also could not be confidently assigned to a known helicase family. The Hachiman-encoded

HamB protein is closely related to SF2 Ski2 helicases, orthologs of which have diverse activities on RNA and DNA substrates (Fig. 3-1B, Supplementary Fig. 3-1)^{151,152}.

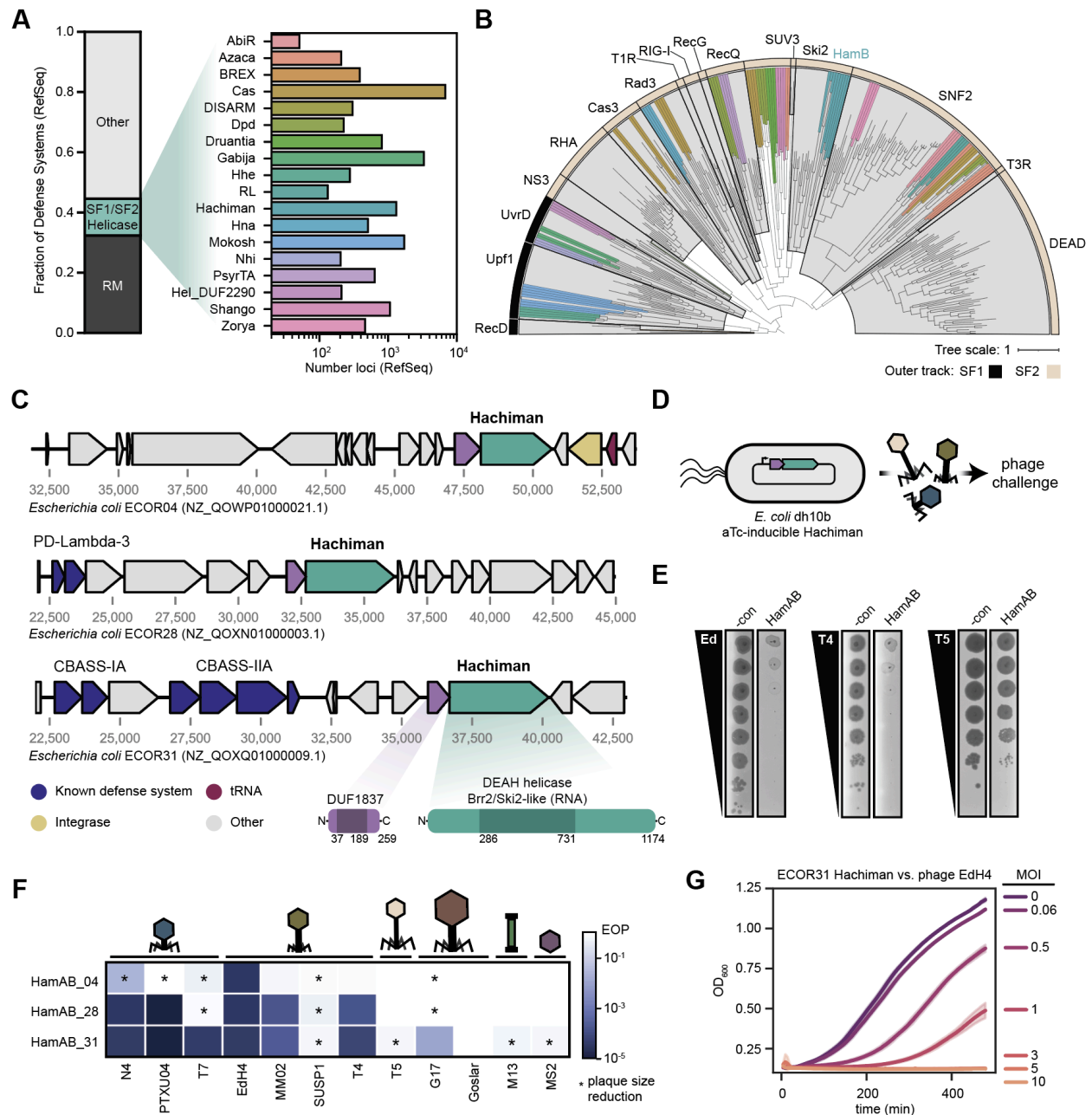


Figure 3-1. Hachiman is a two-component defense system that protects against diverse bacteriophages. **A)** Overview of SF1/SF2 helicase-containing phage defense systems found in RefSeq genomes in the DefenseFinder database¹⁴³. **B)** Phylogenetic tree of core helicase domains of 329 helicases from defense systems from (A) and representative SF1/SF2 helicases¹³⁸. Helicase superfamily is provided in the outer track (SF1 in black, SF2 in tan) and representative families demarcated in gray clades with labels. Defense system-associated helicases are colored as shown in (A). Details on tree construction and sequence alignment provided in methods. **C)** Hachiman loci from *E. coli* strains ECOR04, ECOR28 and ECOR31 tested in this study. HamA genes are shown in purple and HamB genes are shown in green. Additional defense systems identified in PADLOC¹⁵³ are shown in blue, integrases in

yellow and tRNA genes in red. All other genes are shown in gray. **D)** Overview of phage-defense assays. Native Hachiman loci are cloned under an anhydrotetracycline (aTc)-inducible promoter, pTet, and monitored for protection against diverse phages. **E)** Representative plaque assays for ECOR31 HamAB against sensitive phages EdH4 and T4 as well as resistant phage T5. Data are presented as the mean of three biological replicates. **F)** Comparison of different Hachiman loci against 12 diverse phages representing 12 unique phage genera. Data shown represent the mean of three biological replicates. Plaque assays without EOP reductions, but a measurable difference in plaque size are denoted with an asterisk. **G)** Protection against phage EdH4 is complete at low MOI, but insufficient at high MOI. For panels (E-G), Hachiman is induced at 20 nM aTc and for panels (E, F) dCas13d targeting RFP is provided as a negative control. For panel (G), a negative control is shown in **Supplementary Fig. 3-2D**. Data are presented as the mean of three biological replicates $\pm$ standard deviation.

To establish a cell-based assay for assessing Hachiman function, we identified distinct Hachiman loci in *E. coli* strains ECOR04, ECOR28 and ECOR31 using PADLOC (**Fig. 3-1C**, **Supplementary Fig. 3-2A,B**)^{153,154}. We challenged cells expressing Hachiman in plaque assays using *E. coli* phages representing twelve distinct genera (**Fig. 3-1D-E**, **Supplementary Fig. 3-2C**)¹⁵⁵. HamAB from ECOR31 conferred the greatest degree of defense, providing 10^2 - 10^5 -fold reduction in efficiency of plaquing for eight diverse dsDNA phages (**Fig. 3-1F**, **Supplementary Fig. 3-2C**), consistent with the broad-spectrum activity of *Bacillus cereus* Hachiman against *Bacillus subtilis* phages⁴¹. Hachiman conferred near-complete defense against sensitive phages at low multiplicity of infection (MOI<1), but diminished protection at high viral doses (MOI>1) (**Fig. 3-1G**, **Supplementary Fig. 3-2D-F**). We confirmed Hachiman limits the production of new phage particles (**Supplementary Fig. 3-2E,F**). At low MOI, bacterial growth is unaffected, while at high MOI, no growth is observed (**Supplementary Fig. 3-2D**). Nonetheless, phage production remains limited-to-nonexistent during Hachiman-mediated defense (**Supplementary Fig. 3-2E,F**), meaning the interaction between Hachiman and phage leads to cell death and restriction of phage progeny. These data imply Hachiman functions by abortive infection (Abi), a type of programmed cell death in which infected cells sacrifice themselves before phage infection matures, preventing viral spread⁷. To control for artifacts arising from overexpression, we also confirmed that ECOR31 Hachiman defends against phage while under the control of its native promoter and on a low-copy plasmid (**Supplementary Fig. 3-2G**).

The four phages resistant to tested Hachiman systems (T5, MS2, M13 and Goslar) possess unique genome properties. MS2 is an ssRNA phage and lacks a dsDNA genome, while M13 uses rolling-circle replication to produce its ssDNA genome¹⁵⁶. The dsDNA genomes of T5 and Goslar have limited accessibility to defense systems as they are compartmentalized before and during infection, respectively¹⁵⁷⁻¹⁵⁹. Overall, phage challenge experiments suggest Hachiman activity protects against diverse dsDNA phages recognizing and subverting a central feature of dsDNA phage infection.

3.3.2 Structural basis of HamAB complexation

To determine the molecular basis of Hachiman function, we purified HamA and HamB individually. Only HamB was soluble in isolation. Coexpression of the complete native

Hachiman locus produced a complex of HamA and HamB (Supplementary Fig. 3-3A-B). A 2.7 Å cryo-EM structure of the isolated HamAB complex is a 1:1 heterodimer (HamA₁:HamB₁, Fig. 3-2A, Supplementary Fig. 3-3A-G, Supplementary Table 3-1). The domain organization of HamB is generally consistent with Ski2/Brr2 helicases, with two stacked RecA-like helicase domains, RecA1 and RecA2, comprising the helicase core¹⁶⁰. A degenerated winged-helix domain (WH*) and C-terminal α-helical region (CAH, C α-helix) form the likely nucleic acid binding cleft (Fig. 3-2B). At the N-terminus, an α-helical bundle (NAH, N α-helical) common to HamB orthologs but not found in related Ski2 helicases, contributes to binding HamA. At the C-terminus, a barrel-like fold reminiscent of oligonucleotide binding (OB) domains sits on the side of the complex. HamB folds with intact helicase motifs, including active site DEGH residues (Supplementary Fig. 3-3H-I).

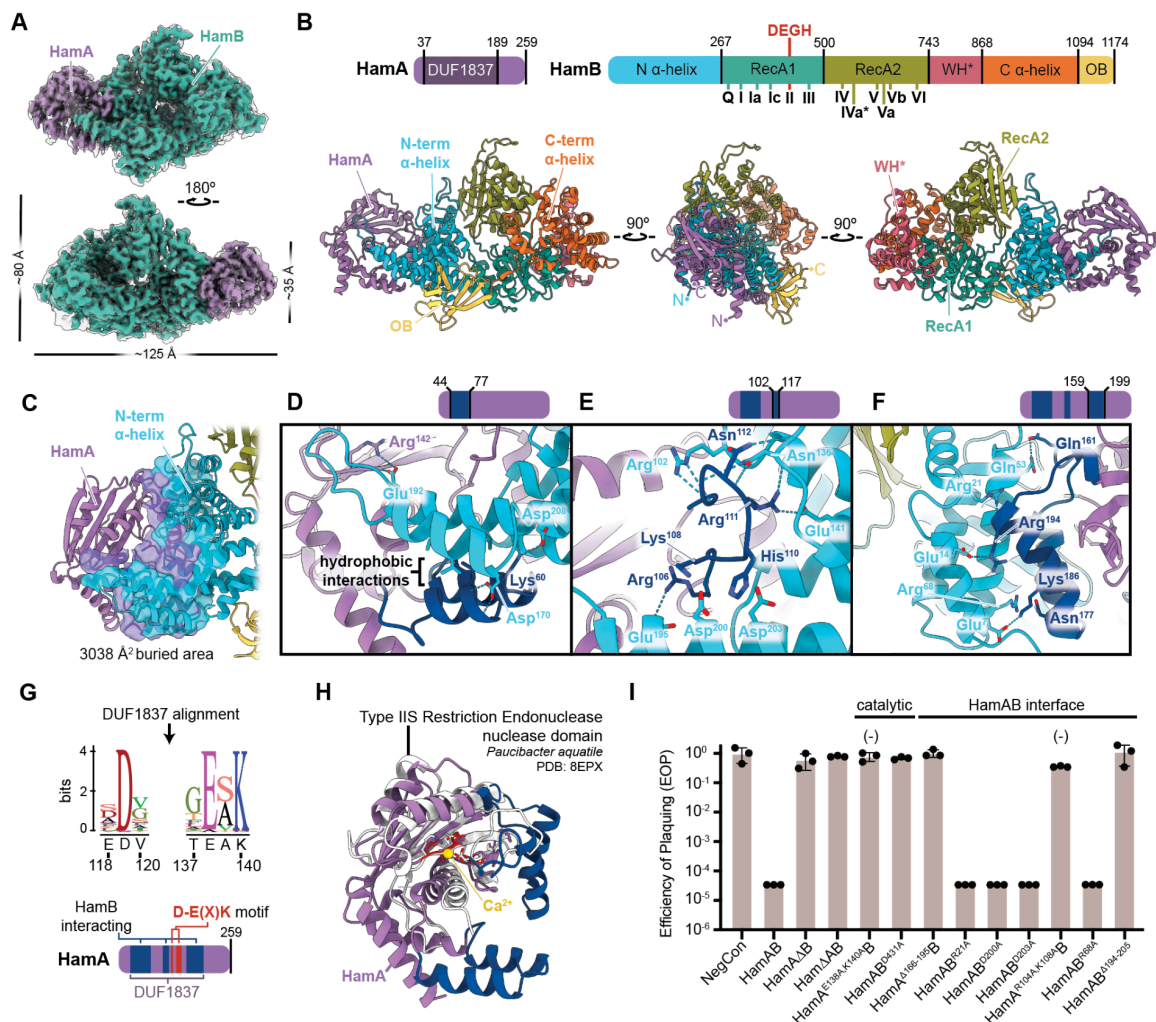


Figure 3-2. Structural basis of Hachiman complexation and identification of the HamA active site.

A) Cryo-EM density of the *E. coli* ECOR31 apo HamAB complex. The sharpened map is colored, while the unsharpened map is overlaid and transparent. **B)** Orthogonal views of the HamAB structure, with domains colored according to the key above. Walker motifs are annotated in the HamB RecA1 and

RecA2 domains. **C)** Overview of the HamA-HamB NAH interface, with surfaces involved in the interaction shown. **D-F)** Detail of three subregions, HamA44-77 (D), HamA102-117 (E), and HamA159-199 (F), contributing to the AB interface. Residues contributing to hydrogen bonding interactions are shown as sticks and are labeled with colors corresponding to the key above each view and in (B). **G)** Sequence logo resulting from alignment of HamA DUF1837 ORFs. The ECOR31 HamA sequence and corresponding positions are shown below each residue logo. **H)** Structural superimposition of the nuclease domain from the *P. aquatilis* type IIS restriction modification system with HamA. **I)** Plaque assays demonstrating the ability of HamAB and various mutants to confer defense against phage EdH4. Individual data points of three independent biological replications are shown along with the mean and standard deviation. The (-) symbol indicates a reduction in plaque size.

The apo HamAB structure shows how HamA contacts the HamB NAH, with three AB interface regions contributing to 3,038 Å² of total buried surface area (**Fig. 3-2C, Supplementary Fig 3-3J**). The first subregion contains a helix-loop-helix which stacks with a HamA helix-loop-helix in the reverse orientation (HamA44-77, **Fig. 3-2D**). The second region is in the center of the interface and includes numerous hydrogen bonds between HamB and an extended HamA-interacting loop (HamA102-117, **Fig. 3-2E**). The third subregion on the HamB N-terminal side is another instance of helical docking with predicted hydrogen bonding and nonpolar interactions (HamA159-199, **Fig. 3-2F**). Structural and biochemical analyses of the AB interface suggest HamB solubilizes HamA (**Fig. 3-2C-F, Supplementary Fig. 3-3A,B**).

3.3.3 HamA DUF1837 encodes a nuclease

The role of HamA is unknown. Alignment of HamA sequences revealed a highly conserved D-E(X)K motif consistent with metal ion-dependent phosphodiester hydrolysis (**Fig. 3-2G**)¹⁶¹. Structurally, HamA is most similar to the Type IIS restriction endonuclease from *Paucibacter aquatilis*¹⁶², with conservation of the core helix/sheet motif (**Fig. 3-2H, Supplementary Fig. 3-3K**). HamA diverges from the *P. aquatilis* nuclease in regions of DUF1837 that facilitate binding to HamB (**Fig. 3-2C-H**).

Cell-based phage defense assays showed that deletion of HamA or HamB, or mutation of their putative active site residues to create HamA(E138A,K140A)B (HamA*B, nuclease-deficient) or HamAB(D431A) (HamAB*, helicase-deficient), ablated defense (**Fig. 3-2I**). Although single interface mutations failed to impact defense, double mutation of a conserved interface motif (R/K)XX(R/K) or deletion of entire helix-loop-helix motifs in HamA blocked Hachiman function. However, these cells were still viable, suggesting HamA must be activated by HamB to trigger effector function. Together, these results show that Hachiman requires both nuclease (HamA) and helicase (HamB) activities for function, and that complexation of HamA and HamB is essential for activation of phage defense.

3.3.4 HamB is a DNA helicase

Despite its functional requirement for phage defense, the identity of the HamB helicase substrate is unclear. Using a malachite green assay that detects orthophosphate

release during NTP hydrolysis¹³⁹, we found strong HamB ATPase activity in the presence of single-stranded DNA (ssDNA, **Fig. 3-3A**). To test HamB DNA helicase activity, we performed DNA unwinding assays by incubating HamB with DNA duplexes of varying lengths and single-stranded overhangs. HamB was capable of ATP-dependent unwinding of a 15 bp duplex with a 15 nt 3' overhang (**Fig. 3-3B**). In addition, HamB unwinds forked, 5' overhang and blunt DNA duplexes, albeit with lower efficiency compared to 3' overhang-containing substrates, suggesting promiscuous substrate acceptance (**Fig. 3-3C-E**, **Supplementary Fig. 3-4A-F**). Longer duplex lengths are not well tolerated (**Fig. 3-3F**). Testing of different DNA duplex and overhang lengths showed that HamB processes a range of DNA substrates, but prefers longer 3' overhangs (**Supplementary Fig. 3-4A-F**).

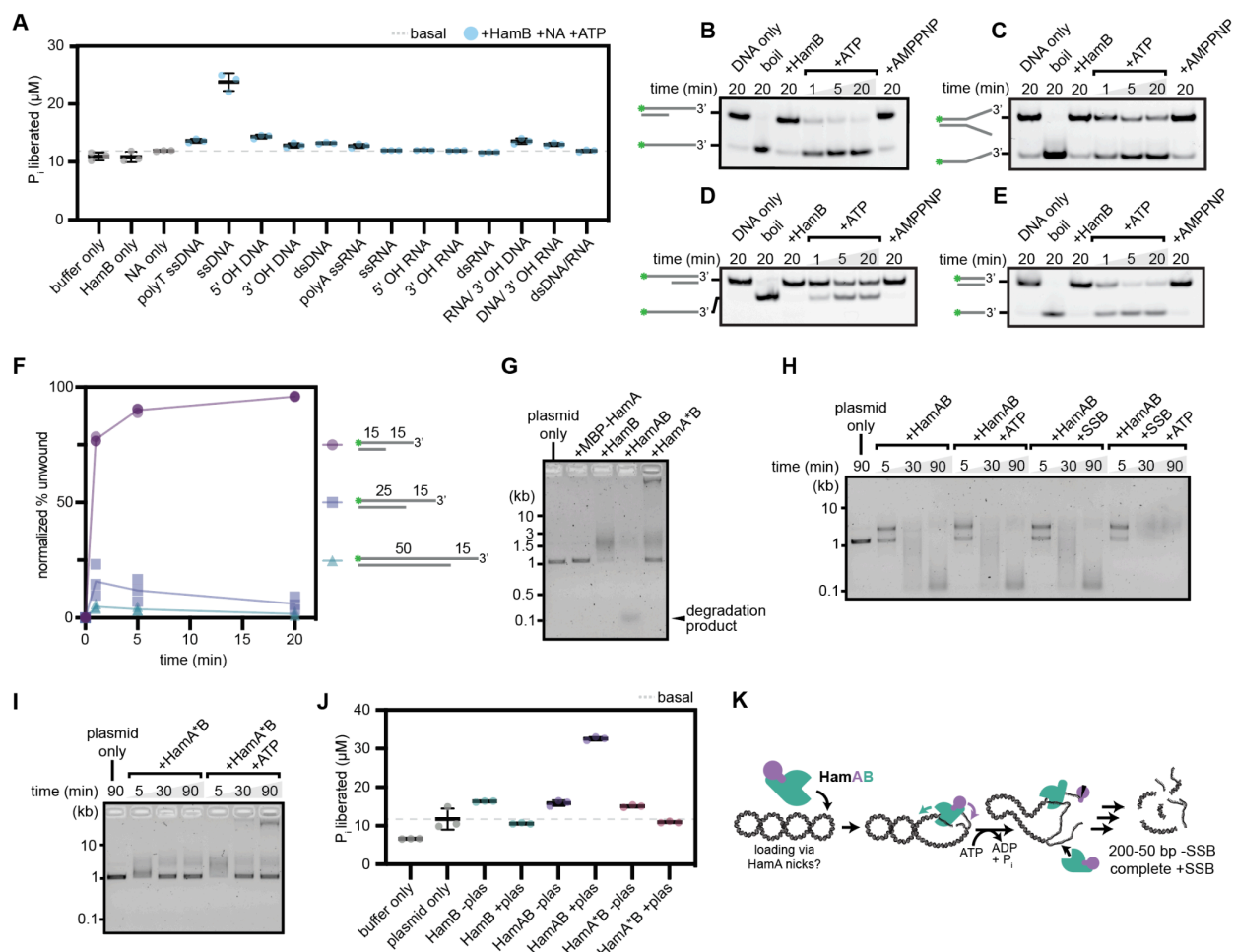


Figure 3-3. HamAB is a DNA nuclease/helicase that degrades plasmids *in vitro*. **A)** Malachite green ATPase assays of HamB against a panel of nucleic acid substrates. Individual data points of three independent biological replicates and the mean and standard deviation are shown. **B-E)** HamB DNA unwinding assays on substrates with a 15 bp duplex and a 15 nt 3' OH (B), forked 15 nt OH (C), 15 nt 5' OH (D), and no overhang (E). DNA substrates are labeled with 5' FAM. Gels are representative of three independent biological replicates. **F)** Normalized percent unwinding of DNA substrates with 15 bp (circles), 25 bp (squares), and 50 bp (triangles) duplex lengths, all labeled with 5' FAM and with a 15 nt 3' overhang. **G)** Plasmid degradation assay. **H-I)** Plasmid degradation assay with SSB. **J)** Plasmid degradation assay with SSB. **K)** Schematic of HamAB mechanism.

OH. Individual data points shown are quantifications of replications of unwinding assays in the format of B-E normalized against basal unwinding (see Methods). **G)** *In vitro* plasmid clearance assay after 90 min at 37°C with ATP using MBP-HamA, HamB, HamAB, and HamA*B visualized on a 0.75% agarose gel. **H)** Time course of HamAB plasmid clearance with addition of ATP or *E. coli* SSB visualized on a native agarose gel. **I)** Time course assay as in H with mutant HamA*B. **J)** ATPase activity of HamB, HamAB, and HamA*B with or without supercoiled plasmid substrates. Individual data points of three independent biological replicates and the mean and standard deviation are shown. **K)** Cartoon depicting a model for HamAB-mediated plasmid degradation.

3.3.5 HamAB degrades plasmids *in vitro*

HamB unwinds DNA substrates, and HamA is a putative nuclease. To determine whether HamA cuts DNA and to ascertain the combined functions of the HamAB complex, we tested activity against purified plasmid DNA. Titration of the wildtype HamAB complex, but not HamA*B, in reactions with supercoiled plasmid DNA show initial plasmid nicking followed by a ladder of degradation products which converged to sizes between 50 and 200 bp irrespective of input plasmid topology (**Fig. 3-3G**, **Supplementary Fig. 3-4G-H**). We also observed cleavage of short dsDNA and ssDNA substrates (**Supplementary Fig. 3-4I-J**). Our observations are consistent with HamA acting as a nuclease effector in Hachiman immunity, though the exact nature of HamA cleavage remains unclear owing to its insolubility *in vitro* (**Supplementary Fig. 3-3A**). Although HamA*B cannot cleave DNA, it forms a low-mobility species upon addition of ATP that may represent a different state captured only when the HamA nuclease is catalytically deactivated (**Fig. 3-3G**).

To assess the possible influence of phage-encoded single-stranded binding (SSB) protein, which has been implicated in activating Hachiman and other antiphage defense systems^{146,163,164}, we incubated reactions with either *E. coli* SSB (EcSSB) or phage T4 gp32 (T4SSB) prior to addition of Hachiman components. Both types of SSB induced complete ATP-dependent plasmid degradation, arguing against direct recognition of phage SSB in Hachiman activation (**Fig. 3-3H**, **Supplementary Fig. 4K-L**). In time-course experiments, addition of either EcSSB or T4SSB accelerated the rate of plasmid interference (**Supplementary Fig. 3-4K-L**). SSB also facilitates complete ATP-dependent degradation of a 75 bp dsDNA, though ssDNA substrates are protected by excess SSB (**Supplementary Fig. 3-4I-J**). EcSSB did not have an observable effect in HamA*B time courses (**Supplementary Fig. 3-4M**), implying HamA nuclease activity is required for subsequent DNA unwinding *in vitro* (**Fig. 3-3I**). SSB stimulates HamAB DNA unwinding and cleavage independent of SSB type, potentially by preventing reannealing of nascent ssDNA (**Fig. 3-3H**)¹⁶⁵.

We noted that the low mobility species observed in HamA*B-plasmid reactions accumulates in an ATP-dependent manner (**Fig. 3-3I**). ATPase assays with HamB, HamAB and HamA*B in the presence of plasmid DNA reveal that unlike HamAB, both HamB and HamA*B suppress ATPase activity upon substrate addition (**Fig. 3-3J**). The

low mobility species may therefore represent a state in which HamA*B ATP-binding enables association with, but not cleavage of, intact dsDNA. These data imply not only that plasmid destruction requires DNA cleavage by HamA, but also that this activity may be coupled with HamB ATPase activity (**Fig. 3-3J**). Considering HamAB does not require ATP to degrade plasmid DNA in the absence of SSB, we postulate that HamAB loads DNA ends induced by HamA nicks generated in vitro by high relative concentrations of complex (**Supplementary Fig. 3-4G**). Nicking triggers ATP hydrolysis which in turn activates further HamA-mediated degradation (**Fig. 3-3K**). When the HamA nuclease is inactivated, HamB does not load DNA ends, but can nonetheless bind intact DNA in an ATP-dependent manner (**Fig. 3-3J**).

3.3.6 Structural basis of HamB DNA binding

Our biochemical studies suggest two modes of DNA binding, one that triggers ATP hydrolysis and cleavage (**Fig. 3-3I**), and one that enables binding of HamAB to intact dsDNA (**Fig. 3-3J**). We observed ATPase activity upon incubation of HamB with a mixed base ssDNA substrate. We performed cryo-EM analysis on a complex of HamB and ssDNA with ATP added during complexation. In the resulting 2.8 Å reconstruction, HamB retains the general domain organization observed in the apo HamAB structure (**Fig. 3-4A-B**, **Supplementary Fig. 3-5A-I**, **Supplementary Table 3-1**), but lacks the extended helix-loop-helix motif that contributes to HamA binding (**Figs. 3-2D**, **3-4B**). Disorder of this region in the HamB-DNA structure further supports its important role in complexation. Surprisingly, despite the addition of only ssDNA to HamB during sample preparation, we observed duplex DNA in the cryo-EM density. The duplex, which is 8 bp in length, arises from a partially palindromic region of the DNA substrate. The duplex appears partially unwound, with one 3' end bound within the HamB nucleic acid binding pocket (**Fig. 3-4C**). Several residues including, but not limited to, canonical Walker motifs contribute to ssDNA binding in the helicase core (**Fig. 3-4D**, **Fig. 3-3H-I**). In the midsection of the duplex, a 'pin' reminiscent of the strand unwinding wedge in the PriA primosomal helicase lies in the center of the duplex and pries the strands apart by pi stacking and physical occlusion (**Supplementary Fig. 3-5J**)^{166,167}. Observation of a 3' end in the entry site of the helicase is consistent with the 3' to 5' polarity determined in vitro (**Fig. 3-3B**). When we mutagenized a conserved threonine which forms a hydrogen bond with the 3' hydroxyl (**Fig. 3-4C**), Hachiman lost antiphage activity (**Supplementary Fig. 3-5K**). Recognition of a 3' DNA end is important for immune activation.

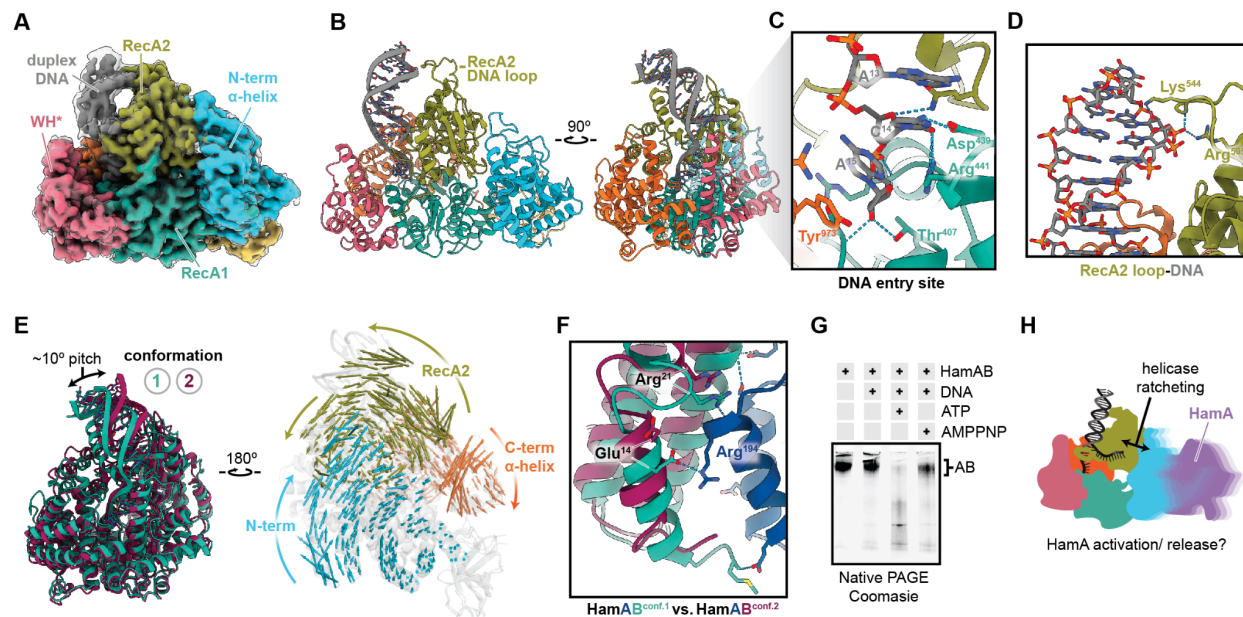


Figure 3-4. Structural basis of HamB-DNA binding and helicase ratcheting. **A)** Cryo-EM density of the 2.8 Å HamB-DNA density. The sharpened map is colored according to domain, while the unsharpened map is overlaid and transparent. **B)** Orthogonal views of the 2.8 Å HamB-DNA structure. **C)** Detail of the 3' end of the DNA buried within the DNA entry site of HamB. Hydrogen bonds and contributing residues are shown with a dashed line. **D)** Detail of the DNA duplex-interacting RecA2 loop. **E)** Left, superimposed conformers of HamB-DNA viewed from the DNA side, with conformation 1 (2.8 Å) colored teal and conformation 2 (2.9 Å) colored burgundy. Right, conformations 1 and 2 viewed from the NAH side and transparent, with vectors colored according to domain representing motion between the two conformations. Vectors are scaled 2x and are calculated using modevectors. **F)** Representative disruption of the predicted AB interface between the two HamB conformations. AB interactions disrupted by HamB motion are shown and labeled. **G)** Native PAGE of reactions of the HamAB complex with the DNA where ratcheting was observed in cryo-EM. ATP and DNA appear to dissociate the AB complex. **H)** Model for HamB signal transduction to the NAH and concomitant release of HamA.

3.3.7 Helicase ratcheting may activate HamAB

We noticed significant conformational variability in the HamB-DNA particle ensemble. Using 3D variability analysis and 3D classifications, we resolved an alternative conformation (conformation 2) of HamB to a nominal resolution of 2.9 Å (**Fig. 3-5A-I**, **Supplementary Table 3-1**, Methods). In the alternate conformation, we observe repositioning of the RecA2, NAH and CAH domains coupled with pitching of the DNA duplex by approximately 10° (**Fig. 3-4E**, left). DNA contacts and the RecA1, WH* and OB folds remain virtually unchanged. Viewed from the HamA direction, the NAH and CAH rotate clockwise, while RecA2 moves counterclockwise (**Fig. 3-4E**, right). Large-scale movement of the NAH-RecA2 interface involves remodeling of interface regions. For example, Tyr525 and Trp521 residues on a distal RecA2 sheet contacting the NAH shift ~9 Å between the two conformations (**Supplementary Fig. 3-5L**). We propose that dynamic switching between HamB conformations represents helicase

'ratcheting' upon entry of the DNA substrate into the active site and triggering of ATPase activity. Motion of the RecA2 domain upon helicase ratcheting transduces to the NAH (**Fig. 3-4E**). The NAH is responsible for binding the HamA nuclease (**Fig. 3-2C-F, 2I**). We superimposed each conformation of HamB bound to DNA with HamB in the apo complex structure. In conformation 1, HamB is in approximately the same position as HamB in the AB complex. Changes in conformation 2 appear to disrupt predicted interactions with HamA (**Fig. 3-4F**), including hydrogen bonds in HamA159-199, a region shown to be essential for defense (**Fig. 3-2I**).

Structural evidence suggests helicase 'ratcheting' transduces motion to the NAH. To test whether helicase motions disrupt the complex, we incubated HamAB with the mixed base ssDNA seen in HamB-DNA structures. Addition of ATP and DNA modulates the HamAB complex, suggesting ATP hydrolysis upon substrate recognition provides input energy to allosterically activate HamAB, potentially by releasing HamA (**Fig. 3-4G**). We observed ATP-dependent disassembly of the complex via size exclusion chromatography (**Supplementary Fig. 3-5M**). Together, these data support a model in which structural changes in HamB may release HamA upon entrance of ssDNA into the helicase active site (**Fig. 3-4H**). In further support of this model, plasmid assays show that nicked DNA triggers ATPase activity, while intact DNA untouched by HamA nucleolytic activity does not activate ATP hydrolysis (**Fig. 3-3J**).

3.3.8 Hachiman degrades phage and host DNA simultaneously during infection

Hachiman binds and degrades DNA. To connect the proposed HamAB structural states with cellular activities, we visualized Hachiman responding to phage infection in vivo using time-course fluorescence microscopy. In uninfected cells, neither wildtype nor inactivated HamAB affected nucleoid morphology (**Fig. 3-5A,B**). When we challenged the control strain lacking Hachiman with sensitive phage EdH4 (**Fig. 3-1F**), we observed that decondensed DNA begins to appear 10 minutes post-infection (mpi, **Fig. 3-5A**). There was no significant difference in the development of this phenotype compared to cells expressing inactive HamAB mutants (**Fig. 3-5B**, HamA*B or HamBA*). However, when we infected cells expressing wildtype HamAB, the nucleoids significantly decreased in size by 30 mpi. By late infection (50 mpi), most cells are 'phantom cells', containing only a small punctum or no visible DNA (**Fig. 3-5B**). Across three biological replicates, at 50 mpi the median DNA cross-sectional area was $<0.2 \mu\text{m}^2$ in our HamAB strain. In contrast, the median DNA cross-sectional area was $>0.9 \mu\text{m}^2$ at 50 mpi in the absence of HamAB or during expression of an inactive mutant.

As determined by time-lapse brightfield microscopy, the average time-to-lysis for EdH4 infecting the control strain under our imaging conditions is ~ 75 mpi (**Fig. 3-5C**). As expected for an abortive infection mechanism (**Fig. 3-1G**, **Supplementary Fig. 3-2D-F**), Hachiman is activated and degrades DNA well before host-cell lysis under wildtype conditions, preventing the release of phage progeny at the expense of cellular viability⁷. These observations are consistent with the abortive infection phenotype

observed in phage production assays (**Fig. 3-1G**, **Supplementary Fig. 3-2D-F**) and agree with biochemical and structural data identifying Hachiman as a DNA-degrading defense system.

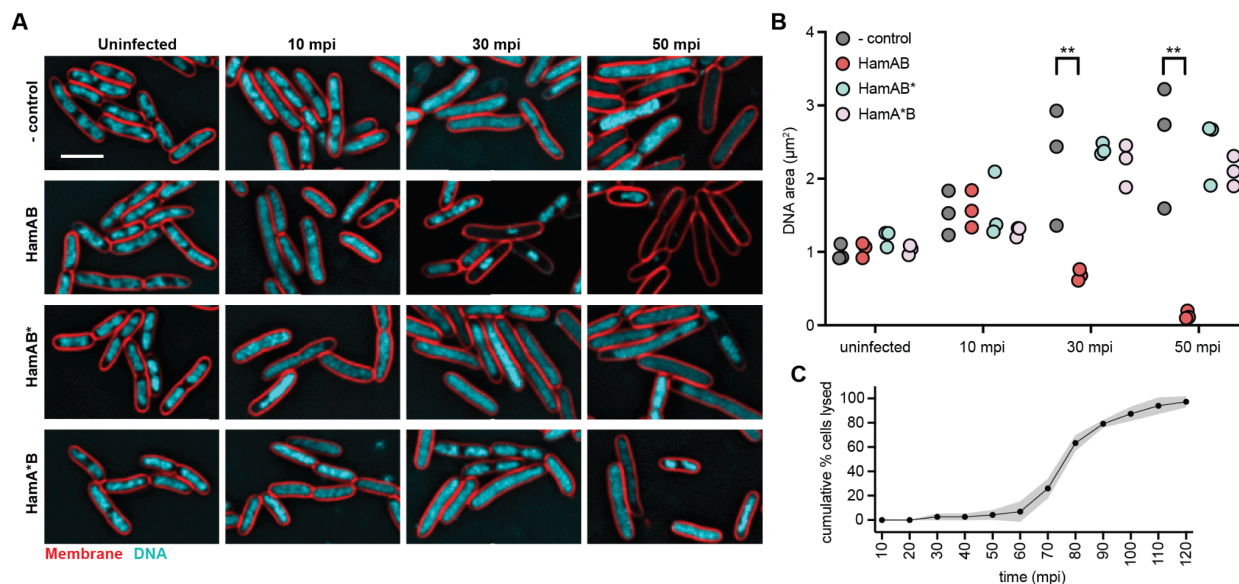


Figure 3-5. Hachiman defends against bacteriophage by nonspecific DNA clearance. **A)** EdH4 infection time-course in *E. coli* expressing wildtype HamAB, HamAB* (HamAB(D431A)) or HamA*B (HamA(E138A,K140A)B) or lacking the Hachiman system (- control). Cell membranes were stained with FM4-64 (red) and DNA was stained with DAPI (cyan). Scale bar = 3 µm. MOI ≈ 2. **B)** Quantification of intracellular DAPI-stained DNA cross-sectional area over the course of EdH4 infection. Dots represent individual medians from three biological replicates. ** p < 0.01 by Dunnett's test. n >225 in total across all replicates for each condition (see Methods). **C)** Time-to-lysis of EdH4 infecting the control strain based on time-lapse brightfield microscopy under the same growth and infection conditions as the time-course in (A) and (B). Black points represent the mean cumulative percentage of total lysed cells that have burst at 10 minute intervals over the course of EdH4 infection, measured in triplicate. Shaded region represents the standard deviation.

3.3.9 DNA damage activates Hachiman

Hachiman responds to phage infection by clearing cells of DNA. The HamB helicase recognizes 3' ssDNA ends and activates ATP hydrolysis (**Fig. 3-2B-F, J**), which in turn activates the HamA nuclease (**Fig. 3-4E-H**). Since Hachiman defends against diverse bacteriophage genera with little or no protein homology, we considered the possibility that Hachiman does not directly recognize a conserved phage component such as phage-encoded SSB¹⁶³. Instead, Hachiman could sense general changes in host physiology.

We reasoned that small molecules that interfere with DNA metabolism might differentially engage Hachiman and elicit bacterial toxicity if it responds to changes in host genome integrity. To test this hypothesis, we treated cells expressing Hachiman

with minimum inhibitory concentrations of DNA-damaging antibiotics. To control for confounding factors arising from potential drug-induced excision of endogenous MGEs, we used the *E. coli* MDS42 strain background, which is devoid of cryptic prophage and insertion sequence elements^{168,169}. We first showed that Hachiman retains antiphage activity invading phages in this strain background (**Supplementary Fig. 3-6A,B**), confirming that Hachiman activity is not dependent on known cryptic prophages and insertion sequence elements. We next treated cells with nalidixic acid (nal), a quinolone inhibitor of DNA gyrase and topoisomerase IV (topo IV)^{170,171}. Aberrant persistence of the protein-DNA linkages during topoisomerase inhibition by nal results in DNA nicks, replication fork arrest, and double-strand breaks (DSBs)^{172–175}. In the absence of bacteriophage, we observed growth inhibition in response to minimum inhibitory amounts of nalidixic acid when WT Hachiman was present compared to HamA- and HamB-inactivated mutants (**Fig. 3-6A, Supplementary Fig. 3-6C**). Novobiocin (novo) is an aminocoumarin that also interferes with gyrase and topo IV, but by an orthogonal mechanism which subverts direct DNA damage^{176,177}. In growth experiments, WT HamAB had minimal differential effects after novo treatment (**Fig. 3-6B, Supplementary Fig. 3-6D**). Rather than inhibit an essential enzyme regulating DNA topology, the polyketide/peptide bleomycin directly induces ss- and dsDNA breaks by generation of radical intermediates¹⁷⁸. Wildtype Hachiman caused elevated toxicity during bleomycin treatment compared to HamA and HamB mutants (**Fig. 3-6C, Supplementary Fig. 3-6D**). Cultures treated with mitomycin C, an alkylating agent which causes interstrand DNA crosslinks and subsequent DSBs, also enhanced toxicity of wildtype Hachiman compared to mutants (**Fig. 3-6D, Supplementary Fig. 3-6E**)¹⁷⁹. In both mitomycin C and bleomycin treatment conditions, we noticed that strains expressing mutant HamB were consistently more sensitive than HamA mutants. As a control, we treated cultures with gentamycin, an aminoglycoside that inhibits translation by binding to the 30S ribosomal subunit¹⁸⁰. Consistent with DNA damage-dependent activation, we observed near-equivalent responses to all Hachiman constructs to gentamycin exposure (**Fig. 3-6E, Supplementary Fig. 3-6F**). Our results demonstrate that Hachiman can reliably be triggered in the absence of bacteriophage, that activation follows direct DNA damage and that activation requires the combined catalytic activities of HamA and HamB.

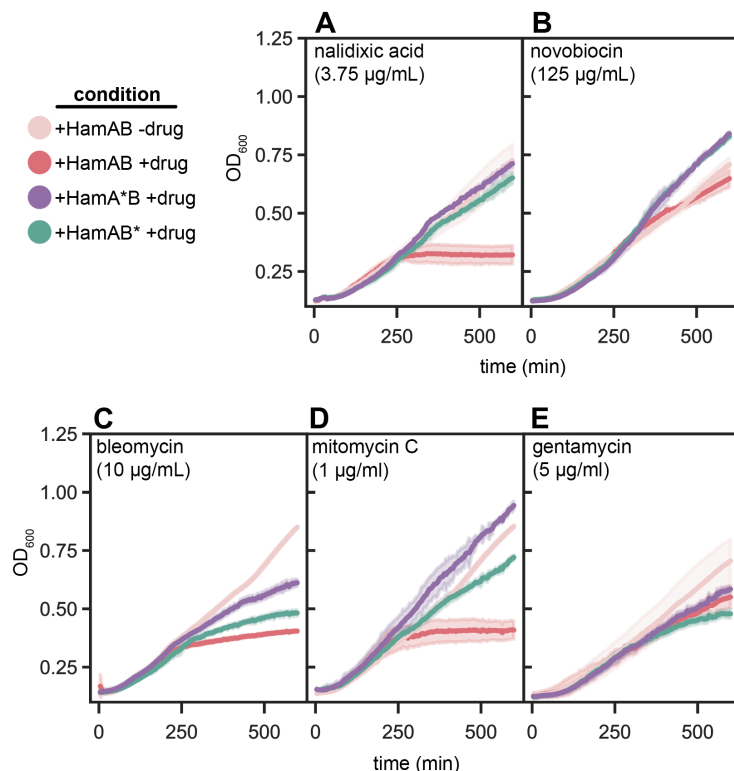


Figure 3-6. DNA damage activates Hachiman. A-E) Cell growth of *E. coli* MDS42 expressing wildtype HamAB, HamA*B and HamAB* at 20 nM aTc in the absence or presence of minimum inhibitory concentrations of nalidixic acid (A), novobiocin (B), bleomycin (C), mitomycin C (D) and gentamycin (F). Growth curves are colored according to condition. See **Supplementary Fig. 3-6** for complete minimum inhibitory concentration determinations. Data are presented as the mean of three biological replicates $\pm$ standard deviation.

3.3.10 Hachiman associates with intact dsDNA

Our data are consistent with a model in which Hachiman triggers abortive infection when 3' ssDNA enters the HamB active site. Although the exact mechanism of DNA damage 'sensing' by HamB remains unclear, we observed formation of ATP-dependent HamA*B-DNA complexes in vitro (**Fig. 3-3I**). We used cryo-EM to visualize this state. HamA*B was incubated with plasmid DNA and ATP for 30 minutes of reaction, after which the specimen was frozen (**Fig. 3-7A**). In the resulting micrographs, many particles can be seen binding intact plasmid DNA (**Fig. 3-7B**). In 2D class averages of plasmid-bound particles, complete HamA*B complexes are seen, with duplex DNA spanning the protein and bending slightly at the point of contact (**Fig. 3-7C**, **Supplementary Fig. 3-7A**). The angle of the DNA in this 'scanning' state is orthogonal to DNA resolved in the 'loading state' in HamB-DNA structures (**Fig. 3-7D**). Masked 3D classification and unbiased alignments produced a map with a 3.2 Å nominal resolution, with lower resolutions (5–7 Å) for the plasmid DNA, although the major and

minor grooves in the central region are apparent (**Fig. 3-7E-F; Supplementary Fig. 3-7A-E; Supplementary Table 3-1**). The dsDNA interacts with the RecA2 loop region (**Fig. 3-7E**). There are few differences between the rest of the complex and the apo HamAB structure (**Fig. 3-2B**). In the molecular model, duplex DNA occupies the same location as RecA2 loop - we could not find an alternate conformation of the loop structure, leading us to conclude that it becomes disordered once DNA is bound (**Fig. 3-7F**). ATP occupies the binding pocket, consistent with biochemical results (**Fig. 3-3I-J**). Contacts made with ATP are in agreement with predictions for HamB Walker and helicase motifs (**Fig. 3-7G; Supplementary Fig. 3-3H-I**). Our observations suggest that HamAB surveys DNA in an alternative 'scanning' mode which may enable monitoring of dsDNA. In this mode, which is facilitated by the RecA2 DNA loop, DNA is restrained from the nuclease active site (**Fig. 3-7H**). Upon phage infection, or conditions which cause elevated levels of DNA damage, HamAB activates by loading of a DNA end, or potentially by recognition of specific DNA structures involved in damage responses such as displacement loops (**Fig. 3-7I**). Entrance of ssDNA into the helicase active site triggers ATPase activity, leading to structural rearrangements enabling HamA activation (**Fig. 3-7J**).

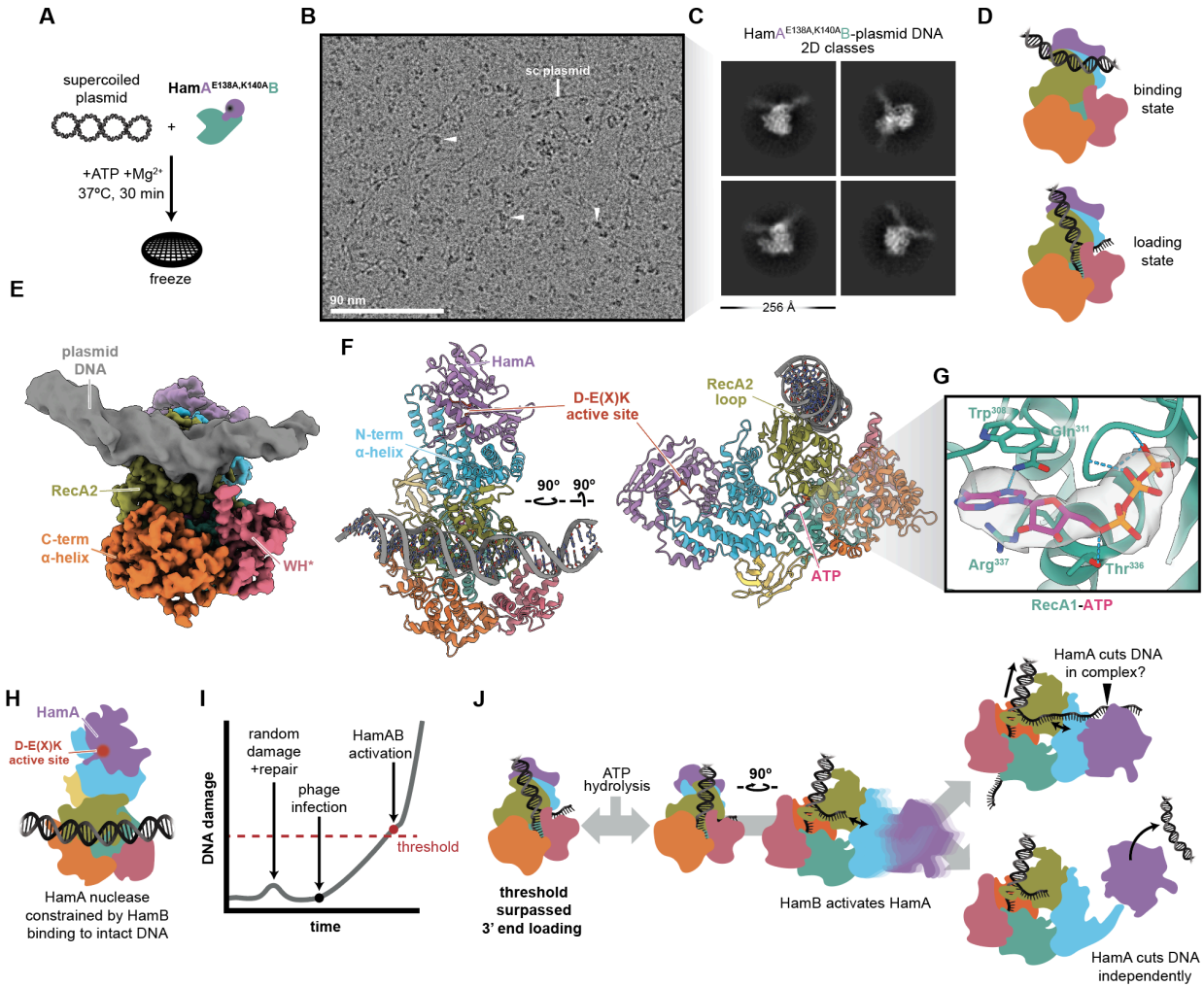


Figure 3-7. Hachiman scans intact dsDNA. **A)** HamA*^B-plasmid +ATP specimen preparation. **B)** Representative motion-corrected, dose-weighted cryo-EM micrograph from the HamA*^B-plasmid DNA dataset. Plasmid DNA and bound particles are indicated with white arrows. **C)** Representative 2D classes of particles bound to plasmid DNA. **D)** Cartoon depicting the scanning state resolved here and comparison with the loading state resolved in the HamB-DNA dataset. **E)** Composite cryo-EM density colored according to domain. Protein regions are from the 3.2 Å deepEMhancer-sharpened map, while DNA is from the 3.2 Å sharp map masked and B-factor refined to display helical features. **F)** Orthogonal views of the HamA*^B-plasmid DNA structure. The DNA sequence is unknown. **G)** Detail of ATP in HamB, with residues and hydrogen bonds shown. The density is masked to ATP. **H)** Cartoon showing separation of intact dsDNA from the HamA active site. **I)** Model of threshold activation of Hachiman. **J)** Proposed mechanism of Hachiman activation.

3.4 Discussion

Our results reveal that the Hachiman prokaryotic defense system is a nuclease-helicase complex, HamAB, that responds to changes in genome integrity. Upon contact with a free ssDNA end, the end inserts into the HamB active site to induce ATP hydrolysis, HamB ratcheting and activation of the HamA nuclease.

Activated Hachiman catalyzes DNA degradation, creating 'phantom' cells cleared of both phage and host DNA, reminiscent of NucC-mediated clearing in some Type III CRISPR-Cas systems¹⁸¹. That Hachiman separates its nuclease and helicase components between two subunits, HamA and HamB, may be intrinsically linked to its robust abortive infection phenotype. Trans nuclease activity from HamA activation could initiate a positive feedback loop. Elevated DNA damage due to HamA activation may then enable other Hachiman complexes to detect new sites of DNA damage, amplifying the immune response. We propose that major changes in genome integrity, such as host genome degradation or recombination-dependent replication by dsDNA phages¹⁵⁶, result in accumulated damage. Extensive damage 'tips the scales' towards Hachiman activation, leading to Hachiman-induced DNA damage and phage restriction through abortive infection. Our results implicate genome integrity as an important battleground during viral infection.

The spread of 'selfish' mobile genetic elements significantly alters nucleic acid metabolism in the cell. For instance, phage T4 degrades the host genome to preferentially replicate and suppress antiphage activities^{182,183}. However, viral teleonomy involves tradeoffs: rapid replication leads to higher error rates¹⁸⁴, increasing the frequency of lesions and stalled replication forks. This is compounded by damage inflicted on the phage genome by other defense systems¹⁸⁵. Phages engage in orthogonal homologous recombination to compensate, a process that universally involves a free 3' ssDNA end and displacement-loop intermediates^{186,187}. We propose that Hachiman senses and activates in response to motifs associated with stress on the integrity of host or phage DNA¹⁸⁸, enabling viral sensing across a range of infection strategies. This explains the broad spectrum protection conferred by Hachiman. The exact identity of 3' end-containing DNA structures which activate Hachiman during phage infection is an area of future interest. We imagine that Hachiman is evolutionarily tuned to activate only when DNA damage is too severe for the host DNA repair machinery to remedy, or if invading entities undergo uncontrolled replication^{156,186}.

A recent study proposed phage-encoded SSB activates *B. cereus* Hachiman¹⁶³. However, our data suggest this mode of activation is a proxy for the true activator of Hachiman encoded in DNA structure. One parsimonious explanation is that expression of phage SSB is incompatible with host replication and recombination machinery while the phage preferentially replicates its own genome. This could result in DNA damage or produce structures such as displacement-loops that appear as DNA damage to Hachiman. In vitro activation of ECOR31 HamAB activity did not require SSB, nor was a differential effect seen when comparing phage and host SSB, consistent with observations in the XPD SF2 helicase¹⁸⁹. Considering the minimal sequence and structural homology between the two stimulatory SSBs, we consider it unlikely that either activates HamB by protein-protein interactions. Our results imply other defense systems believed to be SSB-activated may be stimulated by DNA damage^{164,190}.

As Hachiman both senses and induces DNA damage, Hachiman must be regulated during normal cell activities. We find ectopic Hachiman expression to be mildly toxic (**Supplementary Fig. 3-7F**), in line with prior work¹⁶³, as a fitness cost of carrying a potent and general immune system. During normal cell activities, DNA damage is likely limited and transient thanks to proofreading activities of the native DNA repair machinery (**Fig. 3-7I**)¹⁹¹. How Hachiman interfaces with evolutionarily conserved DNA repair pathways to limit premature activation is of future interest. Cells harboring Hachiman wield broad-spectrum defense, but at the risk of autoimmunity.

We identify HamA, previously DUF1837, as the effector nuclease responsible for DNA clearance. Compared with structural homologs in Type IIS restriction-modification systems, HamA contains insertions that mediate interactions with HamB. Considering the HamA interaction domain (NAH) is present in all HamBs, and that this interaction domain is absent in close relatives in the Ski2 subfamily (**Supplementary Fig. 3-5M**)^{192,193}, we propose HamA insertions were acquired during Hachiman evolution to enable nuclease regulation and allosteric activation. Another distinguishing feature of HamB is the loop on the crown of the RecA2 domain which both facilitates ‘scanning’ of intact dsDNA and forms contacts with DNA during loading into the helicase active site. Comparison to predicted structures of HamB orthologs confirms that either the RecA2 loop or a highly positively charged patch exists at this position, suggesting dsDNA sliding or binding may be a common feature in Hachiman defense (**Supplementary Fig. 3-7G**). Other immune helicases have been proposed to ‘scan’ DNA or RNA for pathogenic signatures^{140,194}. Future studies should address the nature of HamAB DNA surveillance with single molecule techniques.

The AbpAB antiphage defense system was described before the identification of Hachiman¹⁹⁵. AbpAB encodes a nuclease (AbpA) and a Ski2-like helicase (AbpB) with combined activity against DNA. The N-terminal domain of AbpA is similar to the Cap4 endonuclease domain from CBASS systems¹⁹⁶. The C-terminal domain of AbpA is remarkably genetically and structurally similar to HamA, but contains catalytically inactive residues within the HamA active site (**Fig. 3-2, Supplementary Fig. 3-2A, Supplementary Fig. 3-7H**). AbpAB was recently shown to activate in response to mitomycin C-induced DNA damage, in agreement with our data¹⁶⁴. Based on structural and functional homology, we propose AbpAB is a Hachiman variant with an N-terminal fusion in the HamA homolog AbpA (**Supplementary Fig. 3-2A,B**). Why AbpAB would encode a catalytically inactive form of the HamA nuclease while carrying an additional, distinct endonuclease remains enigmatic.

Beyond antiphage immunity, Ski2-like helicases variably accept RNA or DNA substrates¹⁶⁰. Our biochemical results demonstrate HamB is more functionally similar to the Ski2-like DNA helicase Hel308 than Ski2 RNA helicases involved in RNA regulatory processes such as splicing and mRNA decay. Hel308 is conserved in archaea and metazoans, but is absent in bacteria and fungi^{197,198}. Like HamB, Hel308 has a wide

substrate scope, with a preference for 3' to 5' DNA unwinding^{151,193}. Human Hel308 (HELQ) is involved in DNA repair and was shown to localize to sites of DNA damage induced by mutagens in vivo^{199–201}. We observed an analogous response to DNA damage induced by drugs in cells harboring Hachiman. Genetic similarities and functional symmetries between HamB and Hel308 suggest an evolutionary relationship.

This study provides structural and biochemical analyses of Hachiman function that extend our understanding of prokaryotic immune mechanisms. The Hachiman sensor helicase HamB has surprising functional similarities to the archaeal and metazoan DNA repair helicase Hel308, which may explain its general activity against DNA damage. Genome integrity sensing may be a more general role of helicases in immune systems beyond Hachiman.

In this study, we propose a mechanism for Hachiman immunity, but several questions remain. We resolved only partial strand unwinding in HamB-DNA structures, though complete melting was observed for short duplexes in vitro (**Fig. 3-3F**). The role of HamB translocation, and whether it contributes to immunity beyond sensing DNA damage and releasing HamA, is unclear. Relatedly, whether HamB actively 'passes' HamA to loaded DNA (cis cleavage) or simply releases the nuclease for degradation (trans cleavage) is an open question. The relative proportion of cis to trans cleavage is probably salient to the level of defense conferred by Hachiman. Structural views of HamA alone and bound to target DNA could further elucidate the molecular mechanism. Phage-encoded peptide inhibitors of Hachiman were recently reported, though the mechanism of action is mysterious²⁰². Structural data from this work will guide future studies exploring how phages counteract Hachiman defense.

3.5 Methods

3.5.1 Bacterial strains and bacteriophages

For standard cultivation, *E. coli* strains were grown in LB media at 37°C at 250 rpm. Whenever applicable, media was supplemented with carbenicillin (100 µg/mL), chloramphenicol (20 µg/mL) or kanamycin (50 µg/mL) to ensure plasmid maintenance. For bacterial assays, strains were maintained as 25% (v/v) glycerol stocks at -80°C.

Phage propagation was performed using commonly employed protocols. In general, phages were propagated at 37°C in LB Lennox media using an initial MOI of 0.1 and host *E. coli* BW25113 (F- DE(araD-araB)567 lacZ4787(del)::rrnB-3 LAM- rph-1 DE(rhaD-rhaB)568 hsdR514). Phage G17 was propagated on *E. coli* DSM 103255. Phage Goslar was propagated on *E. coli* MC1000. Phages MS2 and M13 were propagated on *E. coli* dh5a F' cells with added 1 mM CaCl₂. All phage titers were determined on their assay hosts harboring a negative control plasmid (pBA635). Infections were carried out as detailed in each section.

3.5.2 Helicase and Hachiman phylogenetic analysis

Proteins chosen for phylogenetic analysis were from DefenseFinder RefSeq db with a RefSeq Protein ID^{58,143}. Proteins from non-Restriction-Modification (RM) defense systems encoding a SF1/SF2 helicase domain were further selected for phylogenetic comparison: AbiR (AbiRc), Azaca (ZacC), BREX (BrxHI, BrxHII), DISARM (DrmA, DrmD), Dpd (DpdE, DpdF, DpdJ), Druantia (DruE), Gabija (GajB), Gao_RL (RL), Hachiman (HamB), Hhe (HheA), Hna (Hna), Mokosh (MkoA, MkoC), Nhi (Nhi), PsyrTA (PsyrT), Rst Helicase+DUF2290 (Helicase), Shango (SngC), Type I CRISPR-Cas (Cas3), Type IV CRISPR-Cas (Csf4/DinG), and Zorya (ZorD). Helicase proteins without RefSeq protein IDs or not tracked in DefenseFinder (ex. Hma) and from RM defense systems (ex. Type I (Type_II_REases) and Type III RM (Type_III_REases)) were not included in the analysis.

For analysis of SF1 and SF2 helicases, 4 randomly chosen examples of the above defense-associated SF1/SF2 helicases were selected and compared to a curated set of SF1/SF2 helicase core domains¹³⁸. To focus our analysis on the helicase core domain, we manually curated the helicase core domain of defense-associated helicases through structural alignment followed by sequence alignment. First, a predicted AlphaFold2 structure of each defense-associated helicase was aligned to the core helicase domain of HamB (PDB ID: 8VXA, this work, residues 286-472, 500-731)^{50,143,203} within the core helicase domain were inferred. Next, for each helicase type, representative helicases were aligned to the corresponding annotated reference using MUSCLE (default parameters), core helicase annotation extracted and insertions removed in Geneious Prime v2023.2.1^{204,205}. To build the helicase tree, sequences were concatenated with a curated set of SF1/SF2 core helicase domains¹³⁸, aligned using ClustalOmega (default parameters), phylogenized using IQ-TREE (-bb 1000, -m MFP (optimal model: LG+R8)), bootstraps inferred using UFBoot2 and visualized using iTOL²⁰⁶⁻²⁰⁹. Clades were inferred by using bootstrap values ≥ 85 followed by analysis of associated sequences.

For analysis of full-length HamA and HamB, proteins were combined and clustered using mmseqs2 (80% coverage, 80% sequence identity)²¹⁰. HamA nucleases and HamB helicases investigated in this study from ECOR04, ECOR28, and ECOR31 and from prior work^{41,164} were spiked into this collection of non-redundant protein sequences for in-group analysis. Sequences were aligned with ClustalOmega (default parameters), phylogenized using IQ-TREE (-bb 1000, -m MFP (optimal model: LG+F+R8 (HamA), LG+F+R10 (HamB))) and visualized using iTOL^{206,207,209}. Clades were inferred by using bootstrap values ≥ 85 in the HamB followed by analysis of spiked-in in-groups.

3.5.3 Identification and visualization of Hachiman-containing loci

Hachiman loci and nearby defense systems were identified in ECOR04 (NZ_QOWP01000021), ECOR28 (NZ_QOXN01000003) and ECOR31 (NZ_QOXQ01000009) using PADLOC¹⁵³. Gene annotations were visualized and represented using dna_features_viewer²¹¹.

3.5.4 Plasmid and strain construction

All new plasmids in this study were constructed through PCR, gel extraction (Zymo D4001) and through Gibson assembly²¹² or Golden Gate assembly²¹³. DNA PCR templates for wildtype and mutant Hachiman loci originated from isolated gDNA (Qiagen DNeasy Blood & Tissue Kit, 69504) from ECOR04, ECOR28 and ECOR31. For most E.coli assays, Hachiman loci were cloned under pTet control into a p15a vector with chloramphenicol (Cm, Sigma) resistance. For E.coli assays involving low-copy vectors, Hachiman loci were cloned under pJEX control into a SC101 vector with Kanamycin (Kan, Sigma) resistance. For protein expression and purification, Hachiman loci were cloned under T7 control in a high copy vector with carbenicillin resistance. In general, plasmids were propagated in dh10b genotype E.coli (F – mcrA Δ (mrr-hsdRMS-mcrBC) endA1 recA1 ϕ 80dlacZ Δ M15 Δ lacX74 araD139 Δ (ara, leu)7697 galU galK rpsL (StrR) nupG λ -) (Intact Genomics). For subsequent phage assays, some plasmids were transferred to E.coli MC1000, E.coli ECOR47 or E.coli dh5a F' where indicated. For protein expression and purification, plasmids were transformed into BL21 AI genotype E.coli (F- ompT hsdSB (rB-mB-) gal dcm araB::T7RNAPtetA). All plasmids used in this study were sequenced-confirmed by full-plasmid sequencing services using Primordium.

3.5.5 Plaque assays

Phage plaque assays were performed using a double agar overlay protocol. Briefly, cultures were grown overnight at 37 °C and 250 rpm. To form overlays, 100 μ L of saturated culture was mixed with molten LB Lennox agar (0.7% w/v agar, 60°C). For assays involving G17 or Goslar a less-dense agar concentration was used (0.35% w/v). The agar-bacterial mixture was supplemented with Cm to a final overlay concentration of 34 μ g/mL and anhydrotetracycline (aTc) (Sigma) concentration of 20 nM. For phages M13 and MS2 an CaCl₂ was added to a final concentration of 1 mM. The top agar and bacterial mixture was poured onto a 5 mL LB Agar and Cm plate and left to dry under microbiological flame for 15 minutes. For plaque assays involving low copy Hachiman loci (strains containing pBA1558 or pBA1747), the agar-bacterial mixture was supplemented with Kan to a final overlay concentration of 50 μ g/mL and no inducers were added. Phages were diluted 10X in SM buffer (Teknova) and 2 μ L of each dilution were spotted onto the top agar and left to dry under microbiological flame. Once dry, plates were incubated at 30°C for 12-16 hours. Plates were scanned in a standard photo scanner and plaque forming units (p.f.u) were enumerated, keeping note of

changes in plaque size relative to a negative control. During assays where “lysis from without”²¹⁴ phenotypes were observed, we interpreted these as a lack of productive phage infection and were approximated as 1 PFU at that concentration. Efficiency of plaquing (EOP) calculations were calculated as $\text{mean}(\text{p.f.u.condition}) / \text{mean}(\text{p.f.u.negativecontrol})$ in Python. The negative control is defined as catalytically deactivated RfxCas13d under pTet control using an RFP-targeting guide²¹⁵. All plaque assays were performed in biological triplicate. Visualizations were performed using GraphPad Prism or Seaborn in Python.

3.5.6 Bacteriophage liquid growth assays and phage production estimation

Liquid phage experiments were performed in a Biotek plate reader using LB + Cm +20 nM aTc media. Strains containing a Hachiman-expressing plasmid (pBA1370) or negative control (pBA1467) were grown overnight at 37 °C and 250 rpm. Strains were seeded into a 96-well microplate reader plate (Corning 3903) at a cfu of ~8e6 cfu per well in 200µL media. For phage experiments, EdH4 was diluted to maximal concentration of 2e10 PFU/mL in assay media, subsequently diluted and 4µL of phage was added to achieve defined MOIs during infection. Growth was monitored in a Biotek Cytation 5 plate reader for 10 hours at 800 rpm shaking at 37°C with OD600 readings every 5 minutes. To estimate free phage particle production at the end of liquid phage assays, wells from infections at defined MOIs were pelleted and the supernatant collected. Phage titers were enumerated via plaque assay on E.coli harboring pBA635 and free phages were determined by dividing by the effective titer at time 0. All liquid phage assays were performed in biological triplicate, sourcing strains from independent overnights. Data were plotted using the matplotlib and seaborn package in Python.

To estimate free phage particle production from a single round of infection, 5 mL cultures were inoculated with ~2e8 cfu of E.coli harboring either a Hachiman-expressing plasmid (pBA1370) or negative control (pBA1467) in LB + Cm +20 nM aTc media. Cultures were incubated at 37°C, 250 rpm for 15 min. Following incubation, ~5e6 pfu of phage EdH4 was added to each culture to achieve a low MOI of ~0.025. Infections were allowed to proceed at 37°C, 250 rpm. At 0, 30, 60 and 90 minutes post infection, 200 µL of infection was sampled, pelleted, supernatant extracted and stored on ice until all samples were collected. Phage titers were enumerated via plaque assay on E.coli harboring pBA635 and free phages were determined by dividing by the effective titer at time 0. Low MOI liquid infection was performed in biological triplicate, sourcing strains from independent overnights. Data were plotted using the matplotlib and seaborn package in Python.

3.5.7 Bacterial liquid growth assays

For Hachiman toxicity-profiling experiments, ig10b strains containing a Hachiman- (pBA1370) or a Hachiman mutant- (pBA1464, pBA1465, pBA1467, pBA1469)

expressing plasmid were grown overnight at 37°C. Strains were seeded into a 96-well microplate reader plate (Corning 3903) at a cfu of ~8e6 cfu per well in 200 µL LB + Cm media. A concentrated stock of aTc was diluted 10X in LB + Cm media and 4 µL added to each well to achieve concentrations of 0, 2, 20 and 200 nM aTc.

For antibiotic sensitivity experiments, an E.coli strain lacking transposable elements including prophages (MDS42) was employed. Strains containing a Hachiman- (pBA1370) or a Hachiman mutant- (pBA1467, pBA1468 or pBA1469) expressing plasmid or a vector control (pBA1801) were grown overnight at 37°C. Strains were seeded into a 96-well microplate reader plate (Corning 3903) at a cfu of ~8e6 cfu per well in 200µL LB + Cm + 20 nM aTc media. Antibiotics (nalidixic acid (Sigma), novobiocin (Sigma), gentamycin (Sigma), mitomycin C (Sigma) and bleomycin (Sigma)) experiments, antibiotics were diluted 2X in LB + Cm media and 4 µL added to each well to achieve final, maximal concentrations of 30 µg/mL, 1000 µg/mL, or 40 µg/mL, 2 µg/mL or 40 µg/mL respectively. Growth was monitored in a Biotek Cytation 5 plate reader for 16 hours at 800 rpm shaking at 37°C with OD600 readings every 5 minutes. Minimum inhibitory concentrations of antibiotic were determined by investigating the lowest concentration of antibiotic that consistently grew to a lower carrying capacity in the vector control than the untreated condition.

All assays were performed in biological triplicate, sourcing strains from independent overnights. Data were plotted using the seaborn package in Python.

3.5.8 DNA Substrate Preparation

Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, IA). Substrates used in unwinding assays were prepared by mixing the fluorescent or larger strand with a 1.5-fold excess of the non-fluorescent strand in hybridization buffer (20 mM Tris-HCl (pH 7.5), 25 mM KCl, 10 mM MgCl₂), and heating to 95°C followed by slow cooling to room temperature for at least an hour. Annealed substrates were purified on an 8% native PAGE gel.

3.5.9 Protein expression and purification

All Hachiman purification constructs were N-terminally tagged with 10xHis-MBP-TEV. For complex purification vectors in the native locus format, only HamA was tagged with 10xHis-MBP-TEV. After transformation into BL21-AI *E. coli*, cells were grown to an optical density of ~0.6 then induced overnight at 16°C with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and 0.1% L-arabinose. Cells were harvested and resuspended in lysis buffer (20 mM HEPES, pH 8, 500 mM NaCl, 10 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA (ethylenediaminetetraacetic acid)-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF), and 10% glycerol). Cells were lysed by sonication, then clarified by centrifugation. The clarified lysate was incubated with

Ni-NTA resin for 1 hr. The resin was washed with wash buffer (20 mM HEPES, pH 8, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP, and 5% glycerol), then bound protein was eluted with wash buffer supplemented with 300 mM imidazole. Eluate was then run over an MBPTrap column (GE Healthcare), washed with MBP/SEC wash buffer (20 mM HEPES, pH 8, 150 mM KCl, 1 mM TCEP, and 5% glycerol), and eluted with MBP/SEC buffer supplemented with 10 mM maltose. Eluted protein from the MBPTrap column was treated with TEV protease overnight. Protease-treated samples were concentrated and run on either a Superdex 200 10/300 GL column (Cytiva) for HamA or HamB solo constructs, or a Superose 6 increase 10/300 (Cytiva) for HamAB complex preparations. Aliquots were snap-frozen in liquid nitrogen for later use.

3.5.10 Cryo-EM sample preparation and data acquisition

The apo HamAB complex sample was rerun over a Superose 6 increase 10/300 (Cytiva) column in Cryo-EM buffer (20 mM HEPES, pH 8, 100 mM KCl, 1 mM TCEP, and 0.5% glycerol). The HamB-DNA complex sample was prepared by combining 15 μ M HamB with 20 μ M DNA in Cryo-EM buffer supplemented with 1 mM ATP and 2 mM $MgCl_2$ and reacting for 30 min at room temperature. Samples were then purified over a Superdex 200 10/300 GL column (Cytiva) in Cryo-EM buffer. The HamA*B-plasmid DNA sample was prepared by combining HamA*B with 1 μ g plasmid in cryo-EM buffer supplemented with 1 mM ATP and 2 mM $MgCl_2$. The reaction was incubated at 37°C for 5 min, after which the sample was frozen. Samples were frozen in liquid ethane using a FEI Vitrobot Mark IV cooled to 8°C at 100% humidity on 2/2 200 mesh UltrAuFoil gold grids (Electron Microscopy Sciences) glow discharged at 15 mA for 25 s (PELCO easyGLOW). In all cases, 4 μ l of each specimen was applied to the grid and immediately blotted for 5 s with a blot force of 8 units.

For apo HamAB and HamA*B-plasmid datasets, micrographs were collected on a Titan Krios G3 equipped with a GATAN K3 Direct Electron Detector in CDS mode and a BIO Quantum energy filter operated at 300 kV and 81,000x nominal magnification in super-resolution mode (0.465 Å/pix). For the HamB-DNA dataset, micrographs were collected on a Talos Arctica equipped with GATAN K3 Direct Electron Detector operated at 200 kV and x36,000 magnification in super-resolution. All cryo-EM data was collected using SerialEM v3.8.7 software²¹⁶. Images were obtained in a series of exposures generated by the microscope stage and beam shifts. For the HamAB apo and HamA*B-plasmid datasets, movies were acquired in an 11x11 pattern. For the HamB-DNA dataset, movies were acquired in a 7x7 pattern.

3.5.12 Cryo-EM data processing

All Cryo-EM data processing was performed in cryoSPARC (v4.2.0 or v4.3.0)¹¹⁶. For the HamAB apo specimen, 4,796 movies were collected and 2 $\times$ binned to a calibrated pixel size of 1.05 Å. 3,314 exposures were accepted after patch motion correction and patch contrast transfer functions (CTF). First, 5,538,550 particles from blob picking

were subjected to 2D classification and ab initio reconstruction of 3 classes, yielding a density consistent with a complete heterodimeric AB complex. The initial ab initio volume was used to create 100 evenly-spaced projection-based templates for further template picking. The 3,505,636 particles from template picking were subjected to 4 class ab initio reconstruction, which gave a HamAB density of 865,990 particles. Further 2D classification and 2D rebalancing (with 9 superclasses) were used to mitigate orientation bias and remove rod-shaped particles missing multiple HamB domains, leading to a final set of 309,630 particles. Single-class ab initio reconstruction followed by non-uniform refinement with on-the-fly defocus and CTF refinement steps gave the final 2.65 Å map¹¹⁸, which was sharpened using DeepEMhancer²¹⁷.

For the HamB-DNA specimen, 9,133 movies were collected and 2× binned to a calibrated pixel size of 1.12 Å. A total of 8,906 exposures were accepted after patch motion correction and patch CTF. Template picker using templates generated from the HamB AF2 prediction gave the best results and were used to isolate 18,254,957 particles at a box size of 256 pix. Reasonable 2D classes were used to train deep picker, which was used to infer 1,125,114 particles at a larger box (512 pix). Ab initio reconstruction followed by non-uniform refinement gave a consensus 2.76 Å density with considerable heterogeneity. Then, 3D Variability Analysis (3DVA) with 3 modes using the ‘simple’ output was used to visualize continuous motion¹⁰⁰. 3D classification with 5 classes was used to resolve densities representing the maxima of motion resolved in 3DVA. Class 1 of the 3D classification gave HamB-DNA conformation 1, which was refined (non-uniform refinement with on-the-fly defocus and CTF optimization) to 2.79 Å and sharpened with DeepEMhancer. Class 0 was refined and sharpened in the same manner, giving HamB-DNA conformation 2 at 2.93 Å.

For the dataset containing HamA*B incubated with plasmid DNA, 3724 movies were corrected for beam-induced motion using patch motion correction, then 2× binned to a calibrated pixel size of 1.05 Å. Contrast transfer function parameters were calculated using patch CTF. Initially, 16,398,369 particles were picked using blob picker from all 3724 micrographs. Multiple rounds of reference-free 2D classification were subsequently performed to remove “bad” particles (i.e., particles in 2D classes with fuzzy or uninterpretable features) yielding 87,084 particles with clear protein characteristics. The particles were then submitted for Topaz training, and the resulting Topaz model was used to pick particles from all 3724 micrographs¹¹⁷, giving a total of 1,322,669 particles. Multiple rounds of reference-free 2D classification were subsequently performed to remove junk particles. After selecting the best classes, 317,540 particles were used for ab initio reconstruction of 3 classes. Of the 3 classes, 2 classes were selected for subsequent heterogeneous refinement. Heterogeneous refinement yielded a good class with 205,538 particles, and non-uniform refinement was performed with the particles from this class, yielding a reconstruction at 2.86 Å resolution. Afterward, multiple rounds of reference-free 2D classification were performed again to select for good particles which presented resolvable features from

2D classification, resulting in 103,451 particles selected. Then, ab initio reconstruction was performed on the selected particles, and subsequently non-uniform refinement, which resulted in a 2.96 Å reconstruction. Then, a focused 3D classification with 4 classes was performed on the predicted DNA binding region of HamA*B, based on views seen in 2D classification, to classify for DNA-bound HamA*B. To generate the focus mask, an atomic model of B-form DNA was built at the predicted DNA binding region, and then a mask of the predicted DNA binding region was artificially simulated using ChimeraX's molmap function with subsequent binarization and softening. The solvent mask was generated to contain both the protein and predicted DNA densities. The best class containing 29,904 particles yielded a classification that was enriched for DNA-bound HamA*B. Then non-uniform refinement was performed on those particles, which resulted in a 3.2 Å reconstruction, which was then sharpened with deepEMhancer. For visualization, a composite map of protein regions from the 3.2 Å deepEMhancer-sharpened map and DNA regions from the 3.2 Å sharp map with B-factor adjustment was made by the color zone segmentation function in ChimeraX.

3.5.13 Model building

The initial model of HamAB was obtained with the ColabFold²⁰³. To build the model, we fit the Colabfold prediction into the experimental HamAB apo density with the fitmap tool in UCSF ChimeraX v1.6.1²¹⁸. There were significant differences in nearly every region of the structure which required iterative manual refinement with a combination of Coot v0.9.4.1¹²⁰, ISOLDE v1.6.0²¹⁹, and Phenix 1.20.1-4487²²⁰. The HamAB apo structure served as the initial model for all other models. The HamB-DNA and HamA*B-plasmid DNA models were built in the manner described above. DNA was built de novo. In the HamA*B-plasmid DNA dataset, the DNA sequence could not be determined, so DNA was modeled as a 31-mer of A-T to maintain base pair interactions during model building. All models were subjected to a final round of Phenix real-space refinement.

3.5.14 NTPase assays

Orthophosphate liberation was determined with a Malachite Green Phosphate Assay kit (BioAssay Systems, Hayward, CA, USA) according to the manufacturer protocol. Briefly, HamB reactions were run in Isothermal Amplification Buffer (henceforth IAB, New England Biolabs, 20 mM Tris-HCl, 10 mM (NH₄)₂SO₄, 50 mM KCl, 2 mM MgSO₄, and 0.1% Tween® 20 (pH 8.8 at 25°C). HamB was diluted to 40 nM, and nucleic acid substrates were diluted to 100 nM, or 4 ng/μL for plasmid reactions, in a total reaction volume of 80 μl in a clear bottom, flat, black 96-well assay plates (Corning Costar). Reactions were allowed to sit for at least 15 min at ambient temperature before initiation with addition of ATP to 1 mM and incubation at 37°C. Reactions were quenched after 30 min with the addition of activated malachite green reagent. The absorbance values of wells were measured after 20 min of color development at ambient temperature with a Biotek plate reader at 620 nm. Orthophosphate liberation

was interpolated against a standard curve with known concentrations of free phosphate. Oligonucleotide substrates are modified from Domgaard et al.¹³⁹

3.5.15 HamAB activity assays

Plasmid interference assays were conducted in IAB. Plasmids were diluted to 4 ng/μL, while other dsDNA and ssDNA substrates were diluted as indicated. Nicked and cut plasmid were generated by treatment with Nt.BspQI (New England Biolabs) and BamHI-HF (New England Biolabs), respectively. After DNA addition, ATP was added to a final concentration of 1 mM where indicated. In cases where *E. coli* (QIAGEN) or phage T4 SSB (gp32, New England Biolabs) were diluted directly from concentrated stock to a final concentration of 400 nM, then the SSB-DNA mixture was allowed to rest for 15 min on ice. Reactions were started with addition of MBP-HamA, HamB, HamAB, or HamA*B to a final concentration of 500 nM, unless otherwise noted, and were incubated at 37°C. Reactions were quenched with addition of EDTA to 10 mM at various time points and were imaged on 0.75% TBE agarose gels in the case of plasmids, or 2% agarose gels for ssDNA and short dsDNA experiments. Gels were stained with SYBR-safe and imaged on a ChemiDoc MP (BioRad).

Complex disassembly size exclusion chromatography experiments were run with elevated concentrations of HamAB (10 μM), ATP or AMPPNP (2 mM) and ssDNA (20 μM) in IAB for a total reaction volume of 100 μL. Buffer was supplemented with KCl for a final concentration of 500 mM. Reactions were incubated at 37°C for 60 min and were then loaded on a Superose 6 increase 10/300 (Cytiva) run with modified SEC buffer (20 mM HEPES, pH 8, 500 mM KCl, 1 mM TCEP, and 5% glycerol).

3.5.16 Gel-shift helicase unwinding assays and quantification

Unwinding reactions were carried out at 30°C in IAB buffer. 100 nM of HamB was incubated with 20 nM DNA substrate for 5 min, and the reactions with protein were initiated by addition of ATP or AMPPNP to a final concentration of 1 mM. At either 1 min, 5 min, or 20 min, reactions were quenched on ice with STOP Buffer (0.4 U proteinase K (New England Biolabs), 18 mM EDTA, 0.36% SDS, and 9% glycerol. Boiled substrates were incubated at 95°C for 5 minutes before immediate loading. Samples were electrophoresed until separation in an 8% TBE polyacrylamide gel at 4°C. Fluorescent bands were imaged using a Typhoon FLA scanner and quantified using Fiji²²¹. The fraction of unwound substrate by HamB was estimated by dividing the intensity of the unwound strand over the sum of the intensities of the unreacted duplex and unwound strand, minus the fraction of unwound substrate from spontaneous unwinding without HamB at 20 minutes, then normalized to the fraction of unreacted duplex without HamB at 20 minutes:

$$\text{normalized fraction unwound by HamB at time } t \approx \frac{\frac{I_{ssDNA,t}}{I_{ssDNA,t} + I_{dsDNA,t}} - \delta}{1 - \delta},$$

where δ is the fraction of $I_{ssDNA,20}$ unwound spontaneously without HamB.

3.5.17 Live single-cell time-lapse and static time-course fluorescence microscopy

Microscopy experiments were performed in biological triplicate. Host cells were grown to OD₆₀₀ 0.3 in LB (+chloramphenicol 30 µg/mL) at 30°C. 12 µL were spotted and spread on the surface of 1% agarose, 25% LB imaging pads containing 30 µg/mL chloramphenicol and 0.05 nM aTc on single-well concavity glass slides, then incubated for 2-2.5 hours at 30°C without coverslips in a humidifier. At this stage, 5 µL of ~2 x 10¹⁰ PFU/mL EdH4 lysate was spotted and spread onto the imaging pads and the pads were incubated at 30°C without coverslips in a humidifier until the desired infection time point. MOI ~2 is estimated based on these initial inocula and infection dose after 2 hrs of incubation before infection and a 30 minute bacterial generation time.

All live cell microscopy was performed on a DeltaVision Elite Deconvolution microscope (Applied Precision, Issaquah, WA, USA). For time-course fluorescence microscopy, imaging pads were stained with 8 µL of dye mix (25 µg/mL DAPI, 3.75 µg/mL FM4-64) and a glass coverslip was placed on top of the pad immediately before imaging at room temperature. For each image, 8 slices in the Z-axis at 0.2 µm increments were collected in each imaging channel (DAPI, FM4-64, brightfield). Exposure times: DAPI = 15 ms, FM4-64 = 15 or 30 ms, brightfield = 8 ms. For time-lapse microscopy, unstained cells were imaged at 5 minute intervals from 5 to 125 mpi at 30°C within the environmental control unit enclosing the microscope stage. 8 slices in the Z-axis at 0.2 µm increments were collected only in the brightfield channel (exposure = 8 ms).

Images were deconvolved in DeltaVision SoftWoRx (version 6.5.2). Image analysis was performed using raw images in FIJI (version 2.3.0/1.53q) and GraphPad Prism (version 10.0.0). Figure panels were created in Adobe Photoshop (21.2.0), GraphPad Prism (version 10.0.0), and Adobe Illustrator (24.2). The following are 'n' values are the total DAPI-stained DNA cross-sectional area measurements for each condition during time-course fluorescence microscopy (with between 46 and 206 measurements per condition per replicate depending on cell density in individual microscopy fields). - control: Uninfected = 525, 10 mpi = 258, 30 mpi = 254, 50 mpi = 249; HamAB: Uninfected = 261, 10 mpi = 229, 30 mpi = 274, 50 mpi = 240; HamAB*: Uninfected = 264, 10 mpi = 272, 30 mpi = 252, 50 mpi = 285; HamA*B: Uninfected = 230, 10 mpi = 255, 30 mpi = 356, 50 mpi = 277. For time-lapse microscopy, the total number of cells lysed across all three replicates = 121, with 37-44 cells per replicate.

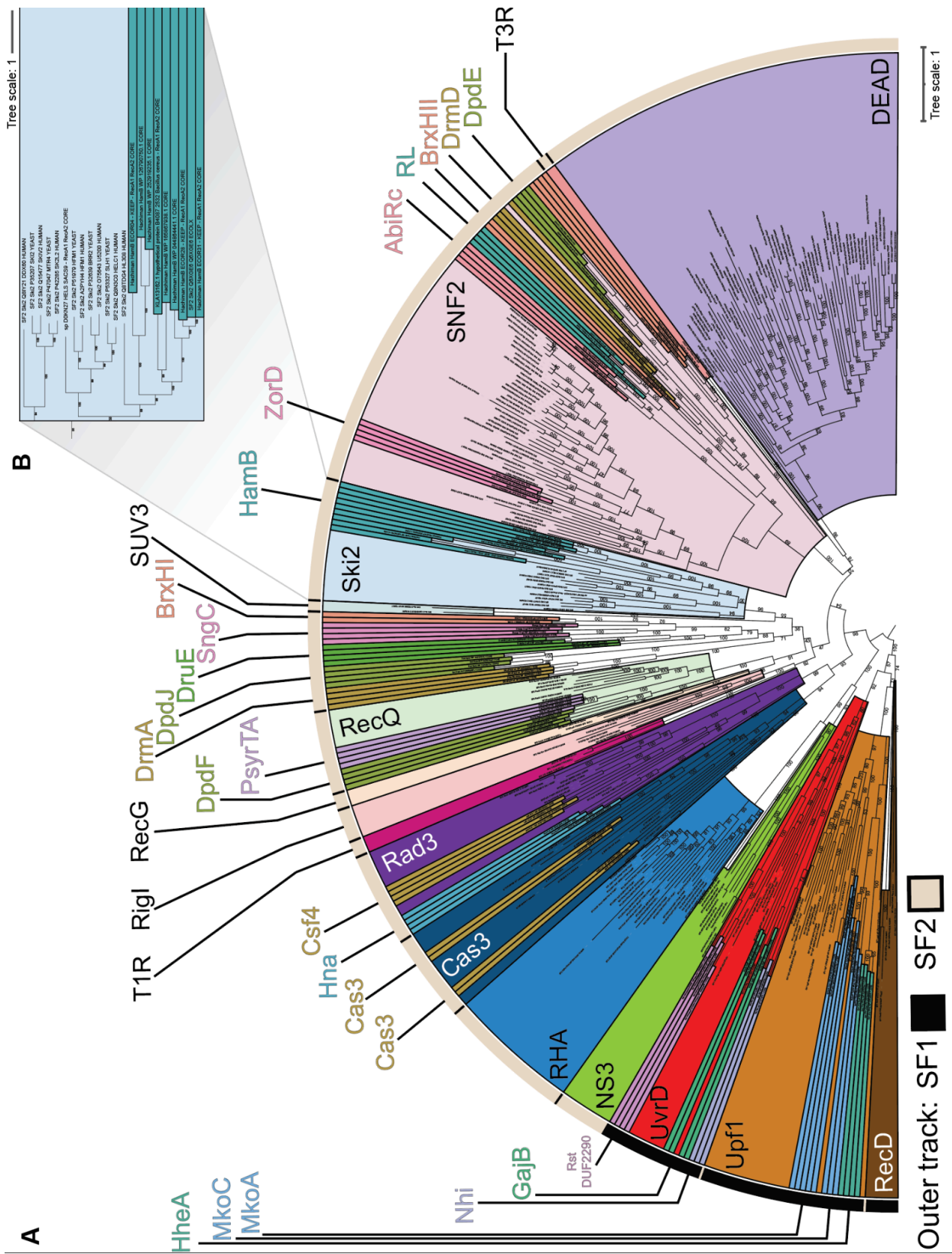
3.5.18 Quantification and statistical analysis

Statistical details for each experiment are found in the figure legend and the accompanying method details. Unless otherwise stated, bar graphs represent the mean of independent biological replicates.

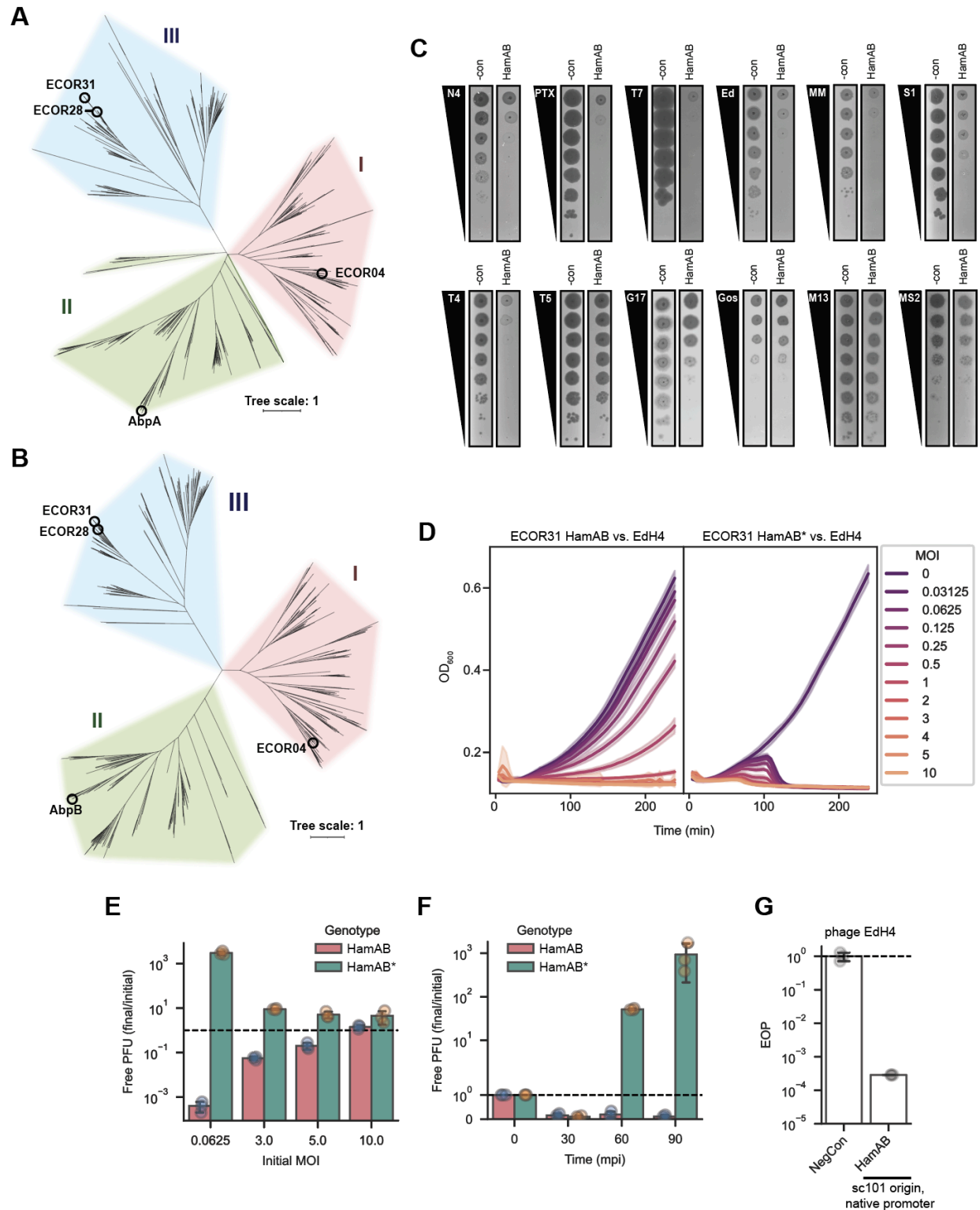
3.5.19 Data and code availability

Structure coordinates and corresponding density maps have been deposited at the Protein Data Bank (PDB) and Electron Microscopy Database (EMD), respectively, under the following accessions: *E. coli* ECOR31 apo HamAB PDB 8VX9, EMD-43613; *E. coli* ECOR31 HamB-DNA (conformation 1) PDB 8VXA, EMD-43615; *E. coli* ECOR31 HamB-DNA (conformation 2) PDB 8VXC, EMD-43616; *E. coli* ECOR31 HamA(E138A,K140A)B-plasmid DNA PDB 8VXY, EMD-43643.

3.6 Supplementary Information

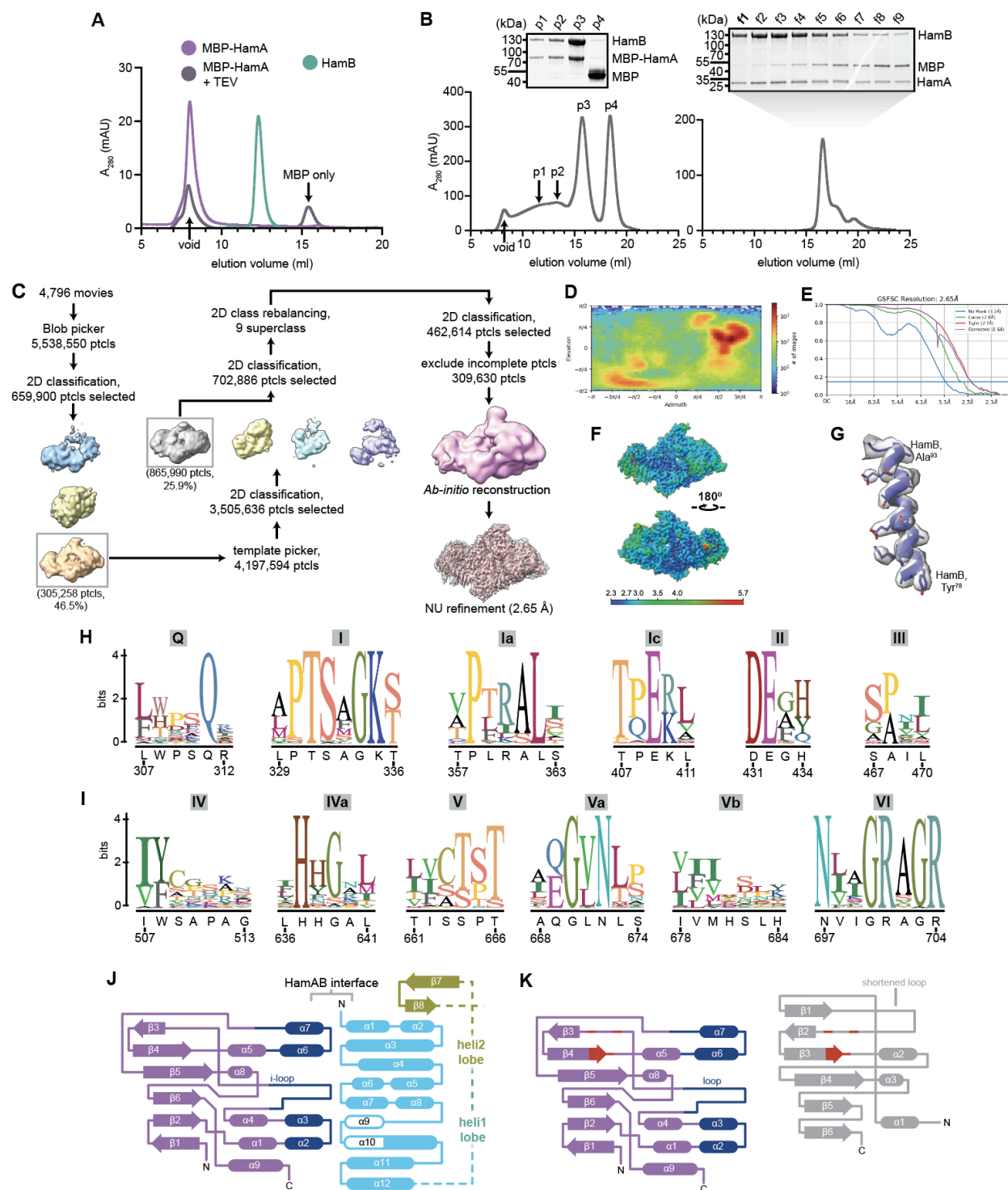


Supplementary Figure 3-1. Helicases and Hachiman phylogenetic analysis. **A)** Annotated phylogenetic tree of phage defense system associated helicase core domains and reference helicases shown in **Fig. 3-1B**. Bootstrapping values determined by UFBoot2²⁰⁸ are shown. **B)** Zoomed-in view focusing on represented Ski2 helicases.



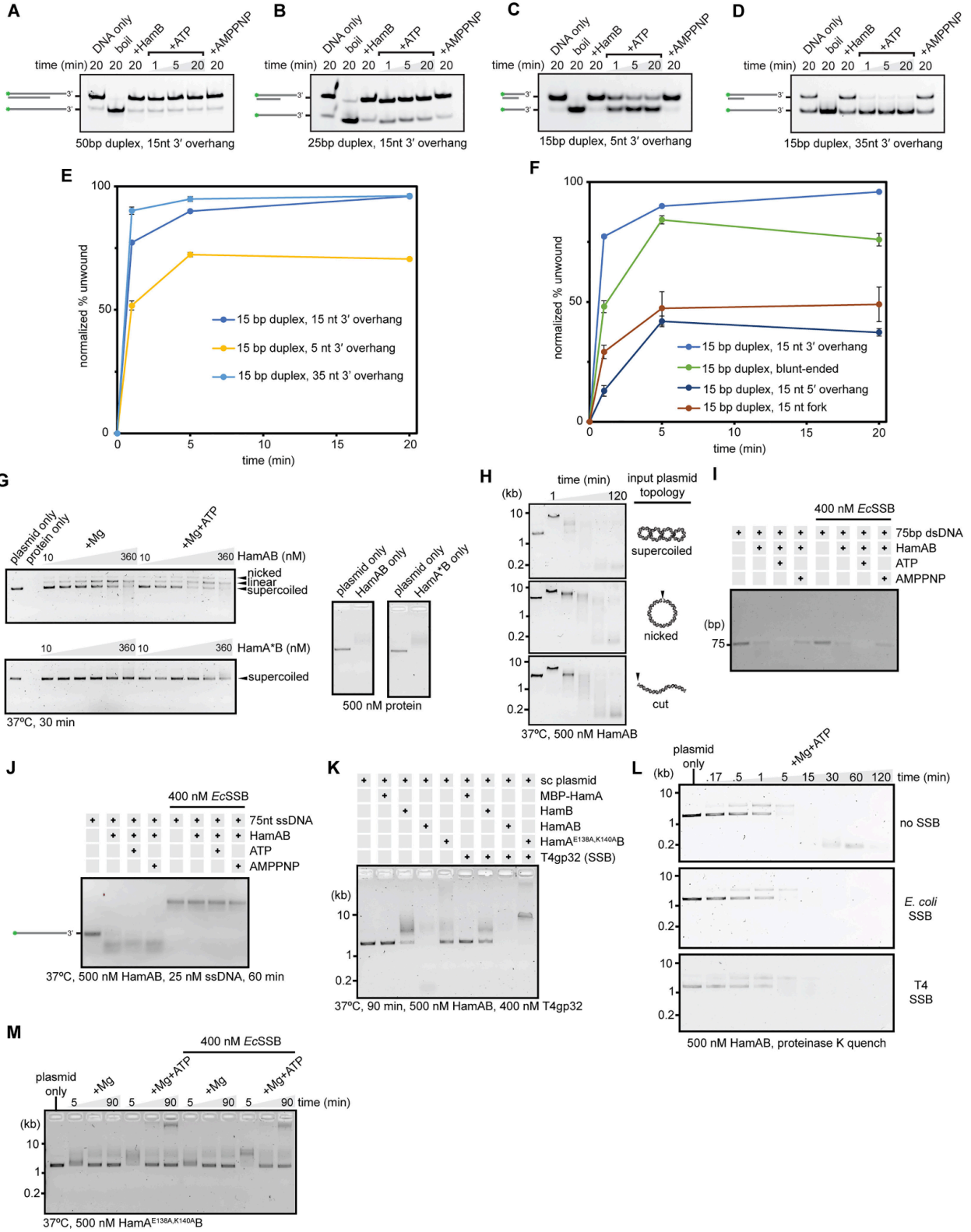
Supplementary Figure 3-2. Function and phylogeny of tested HamAB proteins. A,B) Phylogenetic tree of (A) HamA and (B) HamB from DefenseFinder¹⁴³ with HamA and HamB sequences from this manuscript and AbpAB¹⁶⁴ are labeled. We assign three potential clades of HamB and their corresponding

HamA clades as I-III. **C)** Representative plaque assays for ECOR31 HamAB and phages tested in this study. **D)** Growth curves of E.coli expressing ECOR31 Hachiman (left) or HamB mutant (right) during EdH4 infection at specified MOI. Data are shown as mean $\pm$ standard deviation across three independent, biological replicates. **E)** Phage production assay for EdH4 infection of E.coli expressing wildtype Hachiman (red) or HamB mutant Hachiman (green). EdH4 titers were estimated over time by sampling endpoint supernatants of infections from **Supplementary Fig. 3-2D**. Free PFU was calculated by dividing the infection titer by the phage titer at 0 minutes post infection (mpi). Data are shown as mean $\pm$ std across three independent, biological replicates with individual data points shown. **F)** Phage replication assay for EdH4 infection of E.coli expressing wildtype Hachiman (red) or HamB mutant Hachiman (gray). EdH4 titers were estimated over time by sampling from the supernatant of MOI 0.02 infections. Free PFU was calculated by dividing the infection titer by the phage titer at 0 minutes post infection (mpi). Data are shown as mean $\pm$ std across three independent, biological replicates with individual data points shown. **G)** Efficiency of plaquing for EdH4 infection of ECOR31 Hachiman expressed from a low-copy (sc101) plasmid under its native promoter. Data are shown as mean $\pm$ std across three independent, biological replicates with individual data points shown.

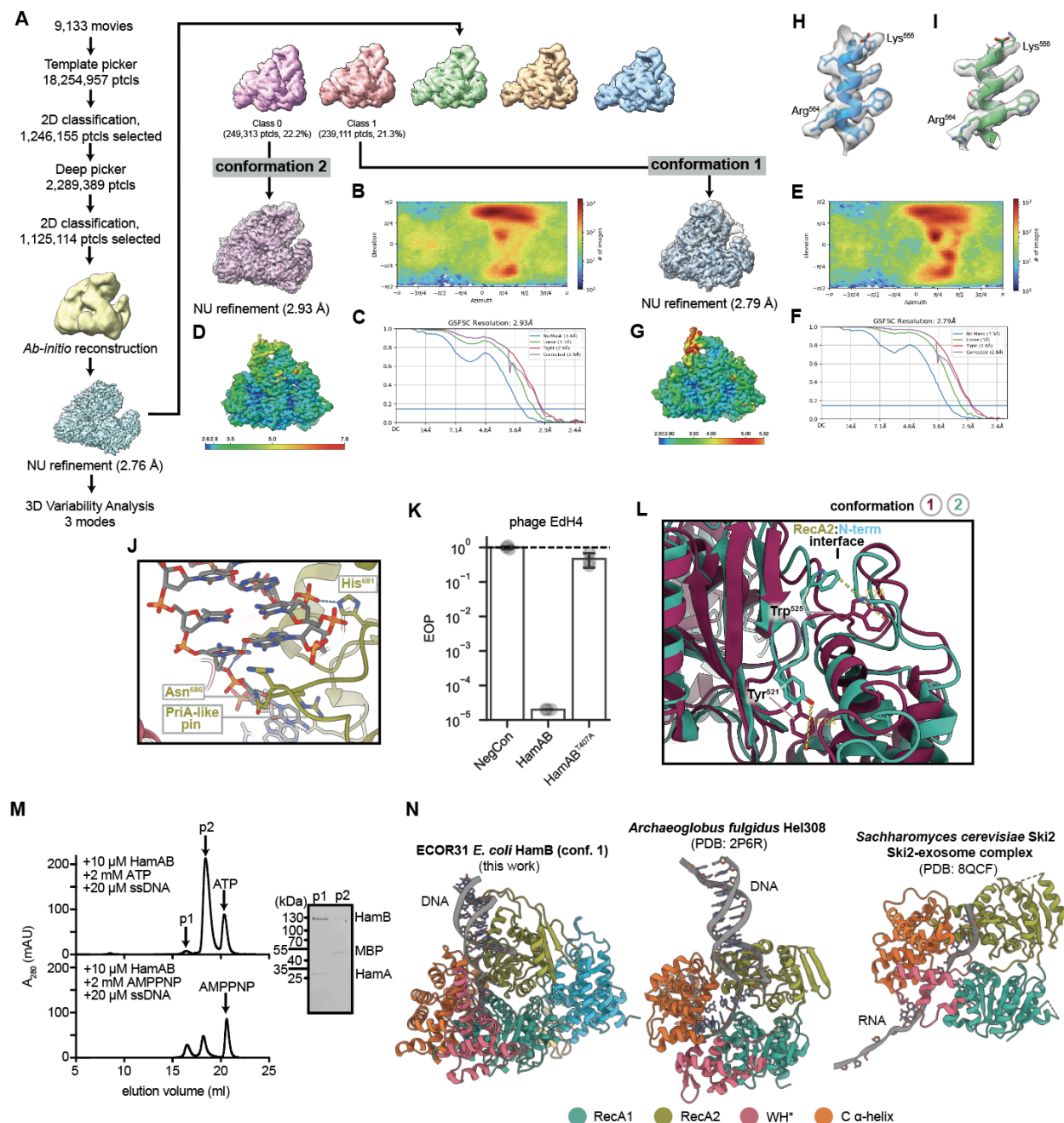


Supplementary Figure 3-3. Cryo-EM structure of the HamAB apo complex. A) Size exclusion chromatography traces of MBP-HamA pre-TEV treatment, post-TEV treatment and HamB alone. **B)** Left, size exclusion chromatography trace of MBP-HamAB and corresponding peaks run on a coomassie PAGE gel. Right, size exclusion chromatography trace of HamAB after TEV protease treatment, with elution fractions run on coomassie PAGE gel shown above. **C)** Particle picking, classification, and refinement strategy to generate the final apo HamAB density. **D)** Orientation distribution of the final

particle set. **E)** Gold standard FSC curve. **F)** Sharpened map colored by local resolution. **G)** Example model-to-map fit. **H-I)** Sequence logos of helicase motifs in the RecA1 (H) and RecA2 (I) domains calculated from the HamB MSA. The residue number and identity of the corresponding sequence in ECOR31 HamB is shown below start and end motif positions. **J)** Secondary structure diagram depicting the HamAB interaction interface. **K)** Comparison of secondary structures of HamA and the *P. aquatilis* Type IIS restriction endonuclease.

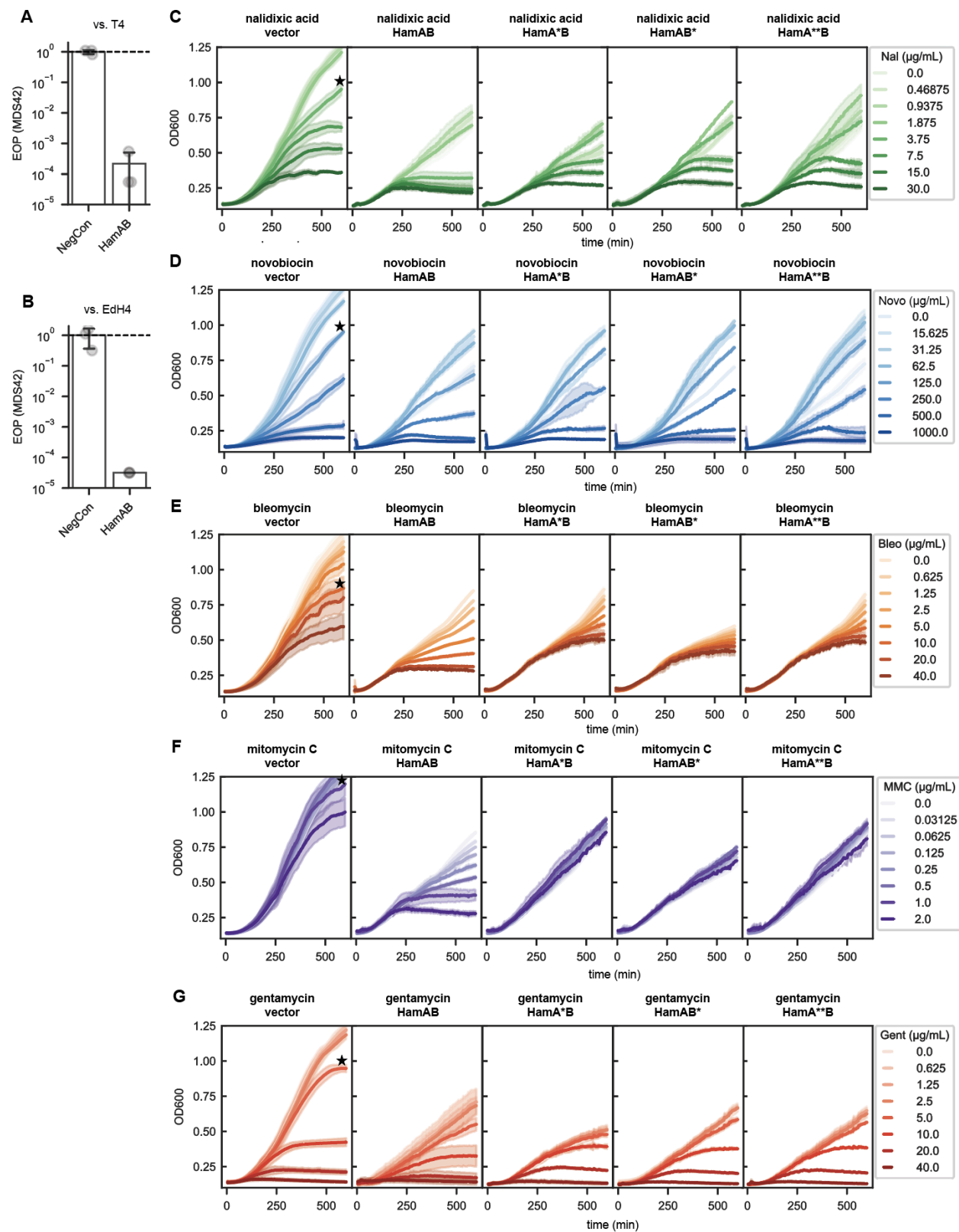


Supplementary Figure 3-4. Analysis of HamAB biochemical activities in vitro. A-D) HamB DNA unwinding assays on substrates with a 50 bp duplex and a 15 nt 3' OH (A), 15 bp duplex and 15 nt 3' OH (B), 15 nt duplex with a 5 nt 3' OH (D), and 15 nt duplex with a 35 nt 3' OH (E). DNA substrates are labeled with 5' FAM. Gels are representative of three independent biological replicates. **E-F)** Normalized percent unwinding of DNA substrates with varying OH lengths (E) and different OHs (F). All substrates are labeled with 5' FAM. Shown are the mean and standard error of quantifications of three replications of unwinding assays in the format of A-D normalized against basal unwinding (see Methods). **G)** Left, titration of HamAB WT and HamA*B complexes with plasmid DNA. Right, Controls of 40 ng plasmid and 500 nM HamAB and HamA*B demonstrating small amounts of contaminating DNA in protein preparations. **H)** Time courses HamAB degradation of supercoiled, nicked, and cut plasmid. **I)** HamAB degradation of a 75 bp dsDNA PCR product with or without EcSSB. **J)** HamAB degradation of a 5' FAM labeled 75 nt ssDNA with or without EcSSB. **K)** Comparison of Hachiman protein and complex plasmid activities with or without phage T4gp32 (T4SSB). **M)** Time course of plasmid degradation by HamA*B with or without ATP and *E. coli* SSB. Conditions in absence of *E. coli* SSB are also shown in **Fig. 3-3H**.

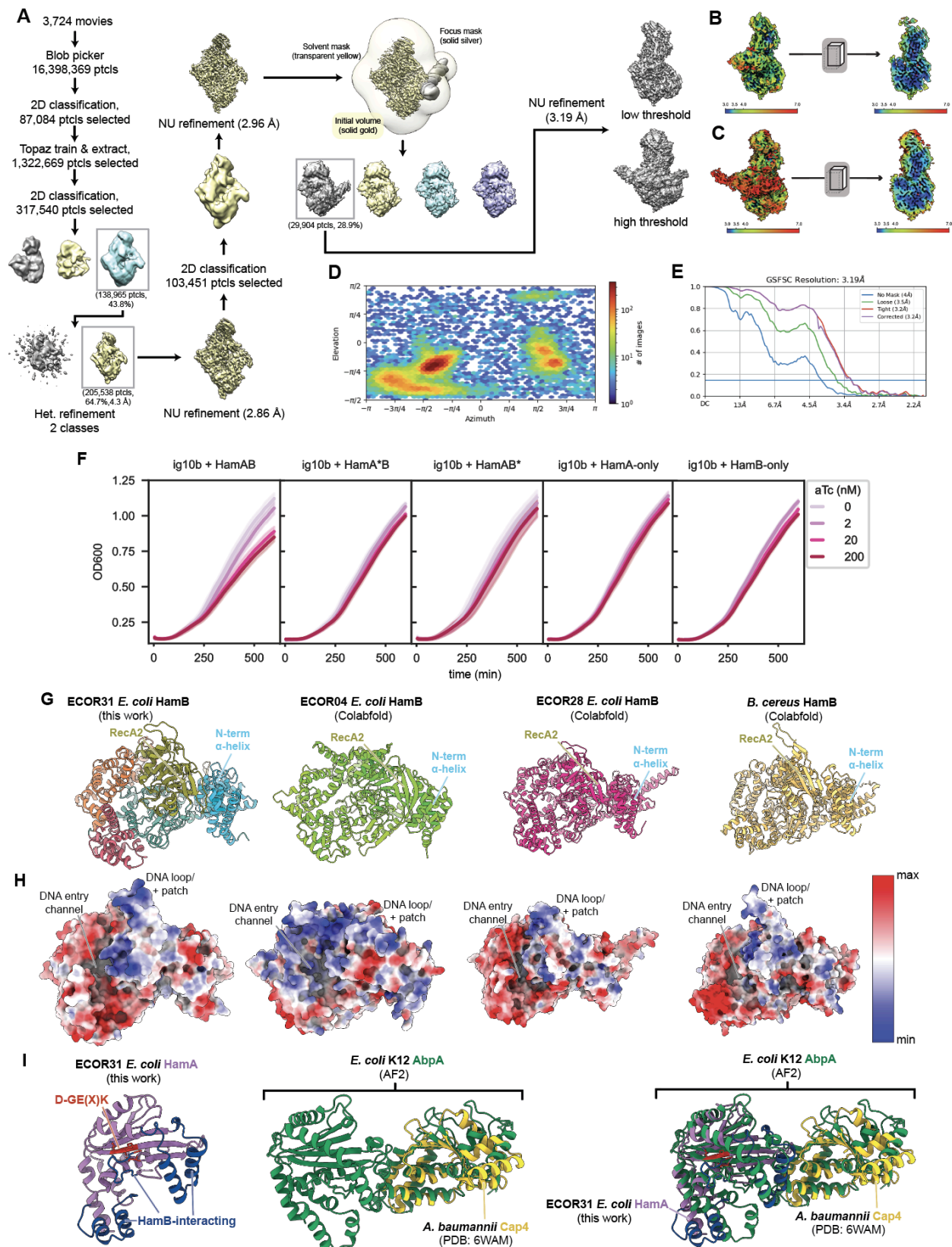


Supplementary Figure 3-5. Cryo-EM structure of two HamB-DNA complex conformations. A) Particle picking, classification, and refinement strategy to generate the final HamB-DNA densities for conformations 1 and 2. **B)** Orientation distribution of the final conformation 2 particle set. **C)** Gold standard FSC curve for conformation 2. **D)** Conformation 2 Sharpened map colored by local resolution. **E)** Orientation distribution of the final conformation 1 particle set. **F)** Gold standard FSC curve for conformation 1. **G)** Conformation 1 Sharpened map colored by local resolution. **H)** Example model-to-map fit for conformation 2. **I)** Example model-to-map fit for conformation 1. **J)** Molecular detail of the PriA-like strand unwinding pin in HamB-DNA conformation 1. **K)** Efficiency of plaquing for a Hachiman mutant (T407A) deficient in the ability to form a hydrogen bond with the 3' hydroxyl of ssDNA. Data are shown as mean $\pm$ std across three independent, biological replicates with individual data points

shown. **L)** Detail of the HamB RecA2-NAH interface and comparison of conformational changes. **M)** SEC traces showing ATP-dependent disassembly of HamAB upon addition of ssDNA. Corresponding coomassie PAGE gel is shown to the right. 10 μ M of HamAB are added in both experiments. ssDNA co-elutes with HamB and MBP. Experiments are run in 500 mM KCl. **N)** Comparison of HamB-DNA with other related helicases bound to their substrates. Middle, *A. flugidus* Hel308 (PDB: 2P6R). Right, Ski2:RNA from a structure of the Ski2-exosome complex (PDB: 8QCF). The exosome complex was hidden for clarity.



Supplementary Figure 3-6. Drug-induced DNA damage activates Hachiman. A,B) Efficiency of plaquing for *E. coli* MDS42 expressing wildtype HamAB or a negative control (dCas13d) against phages T4 (A) and EdH4 (B). Data are shown as mean $\pm$ std across three independent, biological replicates with individual data points shown. **C-G)** Cell growth of *E. coli* MDS42 expressing a vector control (left), wildtype HamAB (second from left), nuclease-deficient HamA*B (middle), helicase-deficient HamAB* (second from right), and HamA^{D119A}B (HamA**B) (right) at 20 nM aTc in the presence of variable concentrations of nalidixic acid (C), novobiocin (D), bleomycin (E), mitomycin C (F) and gentamycin (G). Growth curves are colored according to condition. The minimum inhibitory concentration without the confounding effects of gene expression determined by a vector control used in **Fig. 3-6** is denoted with a star. All growth curves performed in biological triplicate. The mean and standard deviation of each condition are plotted.



Supplementary Figure 3-7. Cryo-EM structure of a HamA(E138A, K140A)B-plasmid DNA complex.

A) Particle picking, classification, and refinement strategy to generate the final HamA(E138A, K140A)B-plasmid DNA density. **B-C)** Final densities at low (B) and high (C) thresholds colored by local resolution. An inner surface slice is shown to the right. **D)** Orientation distribution of the final particle set. **E)** Gold-standard FSC curve. **D)** Conformation 2 sharpened map colored by local resolution. **E)** Orientation distribution of the final conformation 1 particle set. **F)** Toxicity from different levels of Hachiman induction (aTc). All growth curves performed in biological triplicate. The mean and standard deviation of each condition are plotted. **G)** Comparison of ECOR31 HamB from this study with ECOR04 HamB, ECOR28 HamB, and Bacillus cereus HamB Colabfold predictions. **H)** Electrostatic surface potential representations of the structures from F, in the same orientation and scale. The DNA entry channel and RecA2 DNA loop/ positively charged patch are indicated. **I)** HamA and Cap4 structural superimpositions with AbpA.

Supplementary Table 3-1. Cryo-EM data collection, refinement, and validation statistics.

	<i>E. coli</i> ECOR31 apo HamAB (EMD-43613, PDB 8VX9)	<i>E. coli</i> ECOR31 HamB-DNA (conformation 1) (EMD-43615, PDB 8VXA)	<i>E. coli</i> ECOR31 HamB-DNA (conformation 2) (EMD-43616, PDB 8VXC)	<i>E. coli</i> ECOR31 HamA(E138A,K140A)B-plasmid DNA (EMD-43643, PDB 8VXY)
Data collection and processing				
Voltage (kV)	300	200		300
Electron exposure (e-/Å-2)	50	50		50
Defocus range (µm)	-0.8 to -1.8	-0.8 to -1.8		-0.8 to -1.8
Pixel size (Å)	1.05	1.12		1.05
Symmetry imposed	C1	C1		C1
Initial particle images (no.)	5,538,550	18,254,957		16,398,369
Final particle images (no.)	309,630	239,111	249,313	29,904
Map Resolution (Å)	2.65	2.79	2.93	3.19
FSC threshold	0.143	0.143	0.143	0.143

Map resolution range (Å)	2.3-5.7	2.5-4.8	2.6-7.6	2.9-7.0
Refinement				
Initial model used	AF2 model	AF2 model	8VXA (this work)	8VX9 (this work)
Model resolution (Å)	3.08	3.18	3.23	3.87
FSC threshold	0.5	0.5	0.5	0.5
Map-sharpening B-factor (Å²)	-85.9	-92.5	-93.9	-60.9
Model composition				
Non-hydrogen atoms	11,244	9,431	9,391	12,324
Protein residues	1,433	1,139	1,139	1,406
Nucleotide residues	0	24	22	62
Ligands	0	0	0	1
B-factor (Å²)				
Proteins	50.88	53.27	51.88	111.72
Nucleic acid	-	83.34	91.38	314.69
Ligands	-	-	-	89.85
R.M.S. deviations				
Bond lengths (Å)	0.27	0.003	0.003	0.003
Bond angles (°)	0.52	0.530	0.506	0.571
Validation				
MolProbity score	1.87	1.73	1.68	1.99
Clashscore	11.85	10.67	9.14	14.11
Poor rotamers (%)	0.08	0	0	0

Ramachandran plot				
Favored (%)	95.94	96.92	96.83	95.21
Allowed (%)	4.06	3.08	3.17	4.79
Disallowed (%)	0	0	0	0

Chapter 4

Recurrent acquisition of nuclease-protease pairs in antiviral immunity

This chapter is adapted from²²²: Tuck, O.T.*, Hu, J.J.*, Lopez, S.C., Adler, B.A., O'Brien, C.E., Hsieh, K., Meredith, C., Loi, K.J., Yoon, P.H., Doherty, E.E., et al. (2025). Recurrent acquisition of nuclease-protease pairs in antiviral immunity. *Science*, eaea8769.

* these authors contributed equally

O.T.T., J.H.H. and B.A.A. conceived of this project. O.T.T. performed phylogenetic analyses with assistance from J.H.H., K.J.L., P.H.Y. and B.A.A. O.T.T., S.C.L., B.A.A., C.E.O., and K.H. performed molecular cloning and phage challenge assays. O.T.T. and J.J.H. performed protein purification and biochemical experiments with assistance from C.E.O. and C.M. This work was supervised by J.A.D. The manuscript was written by O.T.T., J.H.H., and J.A.D. All authors contributed to reviewing and editing this manuscript.

4.1 Abstract

Antiviral immune systems diversify by integrating new genes into existing pathways, creating new mechanisms of viral resistance. We identified genes encoding a predicted nuclease paired with a trypsin-like protease repeatedly acquired by multiple, otherwise unrelated antiviral immune systems in bacteria. Cell-based and biochemical assays revealed the nuclease is a proenzyme that cleaves DNA only after activation by its partner protease. Two distinct immune systems, Hachiman and AVAST, use the same mechanism of proteolytic activation despite their independent evolutionary origins. Examination of nuclease-protease inheritance patterns identified caspase-nuclease (*canu*) genomic loci that confer antiviral defense in a pathway reminiscent of eukaryotic caspase activation. These results uncover the coordinated activities of pro-nucleases and their activating proteases within different immune systems and show how coevolution enables defense system innovation.

4.2 Introduction

Continual infection by viruses drives the evolution of widespread antiviral immune systems in bacteria and archaea^{41–43,144}. Evolutionarily independent antiviral systems including CRISPR-Cas⁷⁴, CBASS⁶⁰, Aps⁴³, pAgo²²³, retrons²²⁴, Lamassu¹⁴⁴ and Hachiman^{124,225} employ a small, interchangeable pool of genes that are typically fused to core immunity genes to provide new or alternative functionalities^{28,42,61,226–228}. The acquisition of such fusion proteins represents a common but underappreciated mechanism of immune diversification through modular evolution^{28,42,61,226–228}. However, it remains unclear whether newly acquired genes confer similar antiviral activities in the context of distinct antiviral systems, and whether gene acquisition might enable the discovery of new defense pathways^{226,229}.

We identified multiple independent examples of bacterial immune systems with co-occurring genes encoding a trypsin-like protease and a predicted nuclease belonging to the metallo- β -lactamase (MBL)-type hydrolase superfamily^{230–232}. In bacterial antiphage systems, MBL hydrolases resemble nucleases from DNA uptake pathways^{233,234}. While the predicted nuclease is always encoded by a separate, typically upstream gene, the protease occurs as a fusion with different antiviral defense-associated enzymes. First described in AVAST (antiviral ATPase/NTPase of the STAND superfamily, Aps)⁴³, the nuclease-protease gene pair also occurs in Lamassu¹⁴⁴, Hachiman¹⁵³ and DRT (defense-associated reverse transcriptase) antiviral systems²³⁵. How encoded nuclease and protease enzymes contribute to defense in diverse antiviral contexts is unknown.

Here we present evidence for a unified nuclease-protease mechanism common to multiple antiviral systems with different infection-sensing mechanisms in bacteria. Cell-based and biochemical assays show that the nuclease exists initially as an

enzymatically inactive ‘gated’ proenzyme. Site-specific cleavage by the partner protease liberates the nuclease active site. We show that the activated nuclease is specific for DNA, and executes indiscriminate DNA cleavage to restrict viral replication. This mechanism is similar in otherwise unrelated immune systems, suggesting that coordination between the nuclease and protease is conserved in immune systems with different modes of infection sensing. Phylogenetic analysis revealed that gated nuclease homologs additionally exist in genomic loci alongside caspase-family proteases. We found that these caspase-nuclease (canu) loci encode defense systems in which the caspase activates the downstream nuclease, suggesting parallel co-evolution of nucleases with multiple distinct partner proteases. Beyond revealing how paired nuclease and protease enzymes function in multiple antiviral systems, these results demonstrate that modular immune domains can guide the discovery and functional annotation of immune pathways, including systems like canu that display unexpected parallels with eukaryotic caspases.

4.3 Results

4.3.1 Repeated acquisition of paired genes in multiple antiphage systems

We noticed that some Hachiman, AVAST (antiviral ATPase/NTPase of the STAND superfamily, Avs), Lamassu and DRT (defense-associated reverse transcriptase) antiphage defense systems share a pair of co-occurring genes. In each case, a gene encoding a predicted nuclease, belonging to the metallo- β -lactamase (MBL)-fold hydrolase superfamily, sits immediately upstream of a gene encoding a trypsin-like protease fused to the N-terminus of a defense-associated enzyme (i.e. HamA, Avs1a or LmuA) (**Fig. 4-1A**). Across all four systems, the nuclease and protease co-occur and share sequence and structural homology (**Supplementary Fig. 4-1**), suggesting they are evolutionarily and functionally linked.

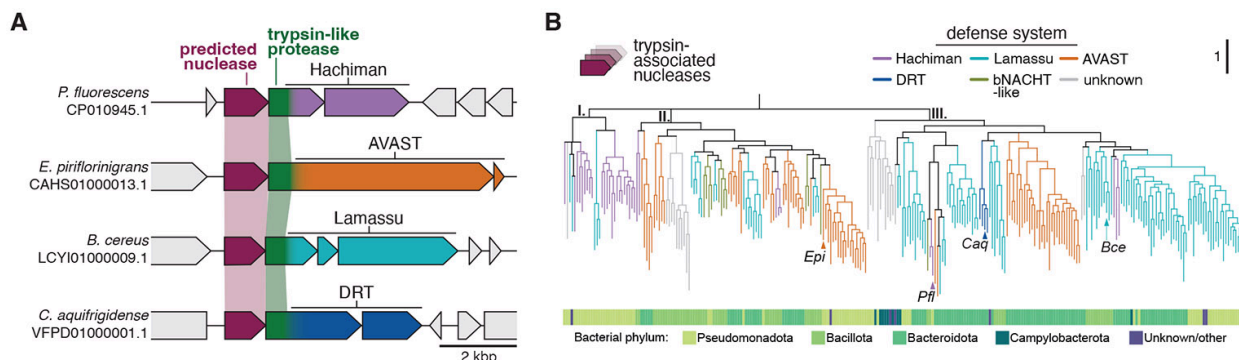


Figure 4-1. Multiple defense systems acquired a predicted nuclease and fused protease. A) Loci illustrating the genetic architecture of multiple defense systems encoding a predicted N-terminal MBL hydrolase-like nuclease and a fused protease. **(B)** Phylogenetic tree of MBL hydrolase-like nucleases accessory to known defense systems which are associated with fused trypsin-like protease domain. Tree

leaves are colored according to inclusion of the nuclease in known defense systems based on neighborhood analysis. The outer track indicates bacterial phyla for each nuclease representative. Leaves shown in (A) are labelled. Branches with support values < 60 are deleted. A subtree is shown. For the complete phylogeny, see **Supplementary Fig. 4-2A,B**.

To determine how antiphage systems acquired nuclease-protease enzyme pairs, we performed sequence- and structure-based searches for homologs of the predicted nuclease, followed by genomic context mining (**Supplementary Fig. 4-2**, see Methods). Our analyses revealed a phylogenetic pattern consistent with frequent exaptation and horizontal genetic exchange (**Fig. 4-1B**). With the exception of rare DRT variants which emerged only once, all systems acquired the nuclease-protease pair on multiple independent occasions. Reconciliation analyses and comparison between nuclease and species trees supports transfer of these genes between defense systems (**Supplementary Fig. 4-3A**, see methods). Moreover, nuclease and trypsin-like protease gene trees are similar, consistent with coevolution of nuclease-protease pairs (**Supplementary Fig. 4-4A**). Our analyses reveal transfer of paired nuclease and protease genes within known defense systems (**Fig. 4-1B**), and identify additional genomic loci with potential immune functions (**Supplementary Fig. 4-2C**).

4.3.2 Hachiman-regulated proteolytic cleavage unmasks the nuclease active site

Our analyses indicate that prokaryotic immune systems frequently acquire nuclease-protease enzyme pairs from other, unrelated systems. To investigate how such acquisition alters immune system function, we began by examining examples of nuclease-protease gene integration into the Hachiman immune system. In ~5% of Hachiman systems, the protease domain is fused to HamA, the nuclease responsible for activation-dependent DNA degradation (**Supplementary Figs. 4-5A, 4-6A**). Analysis of diverse HamA sequences revealed multiple independent examples of protease-HamA fusions, consistent with the corresponding nuclease phylogeny (**Fig. 4-1B; Supplementary Figs. 4-6B-G, 4-7A**). Notably, other types of HamA fusions each occurred just once during Hachiman evolution, but do not co-occur with nuclease genes and are predicted to confer diverse biochemical functions (**Supplementary Fig. 4-6** and **Supplementary Fig. 4-7A-C**)^{124,195,225}. Furthermore, in virtually all of the protease-HamA fusions, the canonical D-GEXK catalytic motif required for HamA nuclease activity is absent (**Supplementary Fig. 4-7D**). This evolutionary pattern implies that fusion Hachiman systems arose independently and lost HamA activity through mutation of active site residues.

We cloned and expressed the *Pseudomonas fluorescens* nuclease-protease Hachiman locus (PflHamMAB) in *E. coli* and verified that it provides defense against phage EdH4 and related phages, reducing viral titer substantially (**Fig. 4-2A,B; Supplementary Fig. 4-8**). Single amino acid substitutions of predicted active site residues in the HamM predicted nuclease, the protease domain of HamA and the HamB helicase each abolished immunity, suggesting that nuclease, protease and helicase activities are all

essential for antiphage activity (**Fig. 4-2B**; **Supplementary Fig. 4-9**). Despite toxicity upon expression, nuclease-protease Hachiman prevented cell lysis at low multiplicity of infection, while high viral doses led to cell culture collapse (**Fig. 4-2C**; **Supplementary Fig. 4-10**). These findings suggest that PflHamMAB, like canonical HamAB systems, defends against phage by inducing programmed cell death or growth arrest.

To determine how nuclease-protease activities enable phage defense, we purified the HamM protein and the HamAB protein complex separately and tested their interactions in vitro (**Supplementary Fig. 4-11**). Incubation of HamM with HamAB resulted in the appearance of three proteolytic fragments and the disappearance of a 42 kDa species corresponding to full length HamM (**Fig. 4-2D**). Mass spectrometry analysis mapped these fragments to specific regions of HamM, with two cleavage sites occurring immediately C-terminal to isoleucine residues I118 and I218 (**Fig. 4-2E**). Proteolysis was absent when we incubated HamAB with other proteins (**Supplementary Fig. 4-12**). These data support the conclusion that the trypsin protease domain of HamA functions as an endopeptidase specific for certain isoleucine residues in HamM. Functional assays further demonstrated that plasmid incubation with either HamM or HamAB alone induced plasmid DNA nicking, whereas incubation with HamM and HamAB together resulted in complete DNA degradation (**Fig. 4-2F**). HamM together with HamAB also degrades short dsDNA, ssDNA, and isolated phage EdH4 DNA, but does not cleave RNA (**Supplementary Fig. 4-13**). These findings support the conclusion that proteolytic cleavage of HamM by protease-HamA converts HamM into a potent deoxyribonuclease.

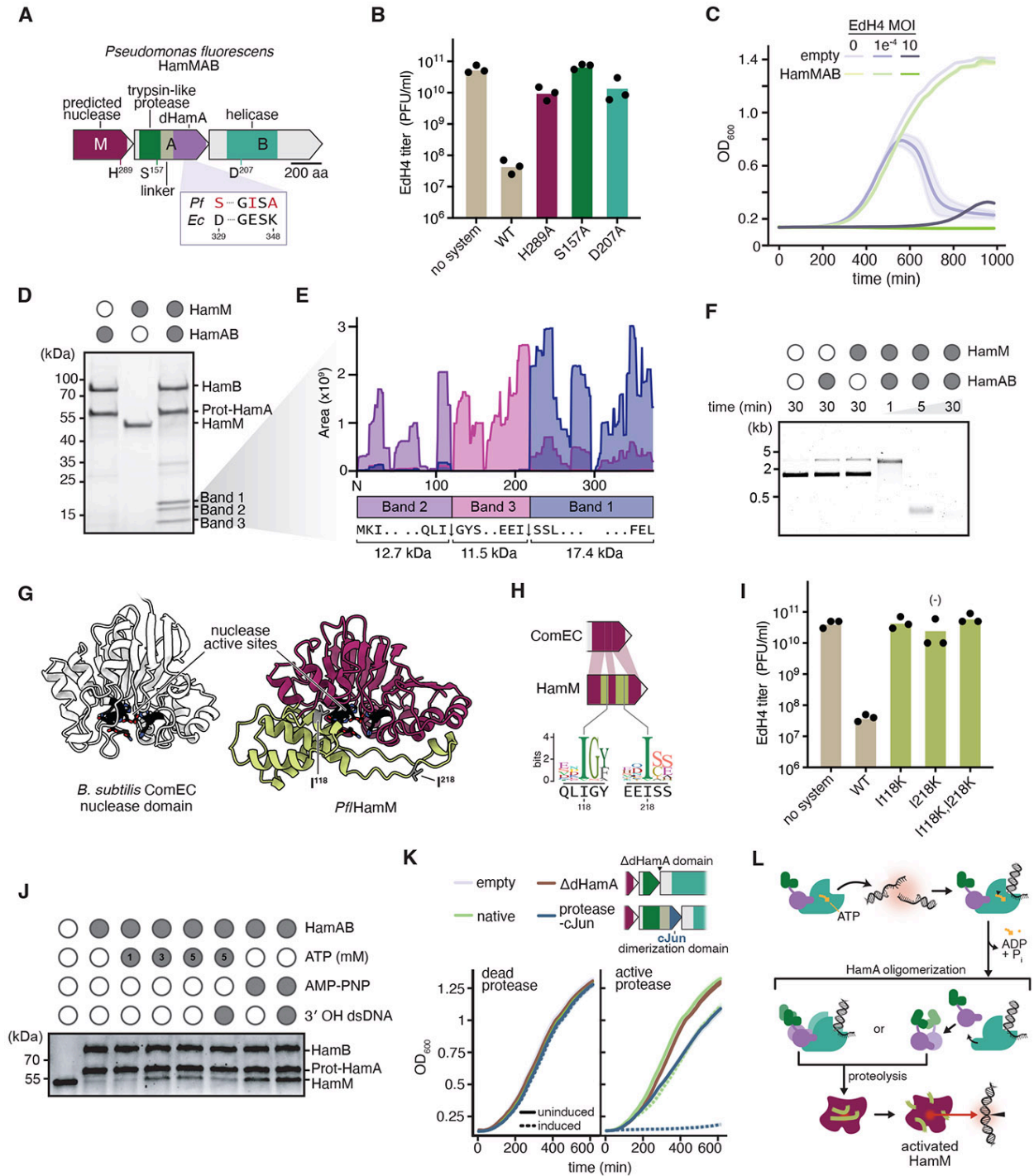


Figure 4-2. HamMAB encodes a specific protease which activates a partner nuclease. A) *Pseudomonas fluorescens* HamMAB operon (CP010945.1) annotated with predicted catalytic residues and comparison to ECOR31 HamA active site. **B)** Quantification of phage EdH4 plaque assays against *Pfl*HamMAB and mutants. Individual data points of three independent biological replicates are shown along with the mean. **C)** Growth curves of *E. coli* expressing *Pfl*HamMAB and empty control during EdH4 infection at specified multiplicity of infection (MOI). Data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates. **D)** *In vitro* reactions of HamM and HamAB

visualized on a Coomassie PAGE gel. **E)** Mass spectrometry reads of fragment 1, fragment 2, and fragment 3 mapped onto the HamM sequence. **F)** In vitro plasmid clearance assay of HamM and HamAB with time course. **G)** Comparison of AlphaFold3 models of the *Bacillus subtilis* ComEC MBL hydrolase-like nuclease domain (uniprot: P39695) with HamM, with active site colored black, protease cut sites colored gray, and insertions colored lime green. **H)** Top: cartoon showing location of insertions in the HamM primary sequence and proteolysis sites. Bottom: sequence logos of cut site position among related nucleases, with the *Pfl*HamM primary sequence as a reference. **I)** Quantification of phage EdH4 plaque assays against *Pfl*HamMAB and protease cut site mutants. Individual data points of three independent biological replicates are shown along with the mean. **J)** In vitro reactions of HamM and HamAB with varying concentrations of ATP, the non-hydrolyzable analog AMP-PNP, and 3' overhang dsDNA visualized on a Coomassie PAGE gel. **K)** Growth curves of *E. coli* expressing variant HamMAB loci either lacking the HamA domain or lacking the HamA domain and fused to the c-Jun dimerization peptide. Left, growth curves in the protease inactive *Pfl*HamMAB background (S157A). Right, growth curves in the wildtype *Pfl*HamMAB background. A solid line represents no induction, while the dashed line represents induction. Data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates. **L)** Model for HamMAB defense, where DNA sensing activates ATP hydrolysis, which leads to oligomerization of protease-HamA and subsequent proteolytic activation of HamM by removal of gating loops.

To understand how proteolysis activates HamM nuclease activity, we compared the AlphaFold-predicted structure of HamM to a homologous MBL hydrolase-type nuclease involved in natural competence^{233,236}. Superposition suggested that HamM includes two disordered insertions relative to the canonical nuclease, both of which appear to block the nuclease active site (**Fig. 4-2G**; **Supplementary Fig. 4-1**). One cleavage-site isoleucine is located on each insertion; both are highly conserved at their respective positions (**Fig. 4-2H**). Hypothesizing that proteolysis removes the active site blockade created by insertions, we made single and double amino acid substitutions of cleavage site residues and tested phage defense. Mutation of either isoleucine residue is sufficient to abolish defense (**Fig. 4-2I**; **Supplementary Fig. 4-9B**). Our results suggest that HamM is a 'gated' pro-nuclease that acquired regulatory insertions responsive to a partner protease. To further confirm that cleavage of insertion regions activates the pro-nuclease, we engineered synthetic breaks into the HamM backbone at identified sites of proteolysis. The resulting 'pre-cut' nuclease is constitutively toxic to host cells, even when its partner protease is catalytically deactivated (**Supplementary Fig. 4-14**).

4.3.3 DNA sensing and oligomerization trigger HamMAB defense

HamM is a pro-nuclease activated by a protease fused to HamA, but the mechanism of protease-HamA regulation remains unclear. We reasoned that DNA damage or DNA intermediates generated during phage replication triggers HamMAB immune activity, akin to canonical Hachiman systems. To test this hypothesis, we treated cells expressing HamMAB with subinhibitory concentrations of nalidixic acid, a DNA-damaging antibiotic. We observed a mild growth defect suggestive of genome integrity sensing, in line with prior work on Hachiman systems (**Supplementary Fig. 4-15**)^{124,164}. To gain molecular insight into immune regulation and activation, we tested

the effect of different ATP concentrations on HamM proteolysis. Consistent with evidence from canonical and fusion Hachiman systems^{124,237}, we found that ATP inhibits immune activation: upon exposure to DNA substrates, ATP hydrolysis led to increased HamM activation (**Fig. 4-2J**; **Supplementary Fig. 4-16**). In support of the role for ATP hydrolysis in Hachiman activation, treatment with a non-hydrolyzable ATP analog (AMP-PNP) stalled HamM processing even when activating DNA was present (**Fig. 4-2J**).

Given that HamAB activates upon DNA sensing and ATP hydrolysis, we sought to determine the role of the dHamA domain in triggering HamM. HamA forms a complex with HamB and therefore likely controls its fused protease^{124,225,237}. We removed the dHamA domain and tested for growth defects in cells. Surprisingly, no toxicity was observed, suggesting an alternative role for HamA during activation (**Fig. 4-2K**). Because oligomerization was shown to activate fused effectors in related systems^{44,238,239}, we examined whether dHamA induces protease oligomerization by replacing it with the leucine zipper dimerization domain from the human c-Jun protein^{238,240}. This protease-cJun construct results in potent toxicity, while catalytically deactivated protease-cJun-expressing cells are healthy (**Fig. 4-2K**). These data suggest the protease alone is insufficient for nuclease activation, and that HamA-mediated oligomerization is necessary for proteolysis. In summary, we propose that upon DNA sensing, HamB hydrolyzes ATP and triggers protease-HamA, leading to HamA oligomerization and proteolytic cleavage of HamM, unmasking the nuclease active site and unleashing nuclease activity (**Fig. 4-2L**)^{124,225,237}.

4.3.4 Gated nuclease activation is conserved in AVAST

We wondered whether nuclease-protease enzyme partners retain functional properties despite their presence in divergent defense systems. To test this possibility, we investigated the nuclease-protease pair in an AVAST antiphage system from *Erwinia piriflorinigrans* (*EpiAvs*, **Fig. 4-1A**), a phage-protein sensing antiphage system with otherwise no homology to Hachiman^{43,44}. We confirmed potent defense against phage P1 and specific activation in response to expression of viral terminase, consistent with prior work (**Fig. 4-3A,B**; **Supplementary Fig. 4-17**). Structural prediction of the AVAST nuclease homolog (*EpiAvs1a*) shows dual insertions that appear to block the conserved active site, including two conserved isoleucine residues, analogous to structural features of PflHamM (**Fig. 4-3A**). We constructed single and double amino acid substitutions of conserved isoleucine residues and tested for AVAST-mediated defense against phage P1 in *E. coli*⁴³. Each mutation alone reduced defense, while the double mutation eliminated AVAST-mediated protection entirely (**Fig. 4-3B**; **Supplementary Fig. 4-18**). To confirm that nuclease is cleaved by the Avs1b protease during infection, we constructed N-terminally tagged versions of *EpiAvs1a* and monitored cleavage by Western blot. A cleavage product corresponding with proteolysis adjacent to isoleucine 134 appeared upon infection, but was absent in cut site mutants (**Fig. 4-3C**). These

data suggest that cleavage of a single gating loop is sufficient to activate the pro-nuclease, consistent with phage defense data (**Fig. 4-3B, C**).

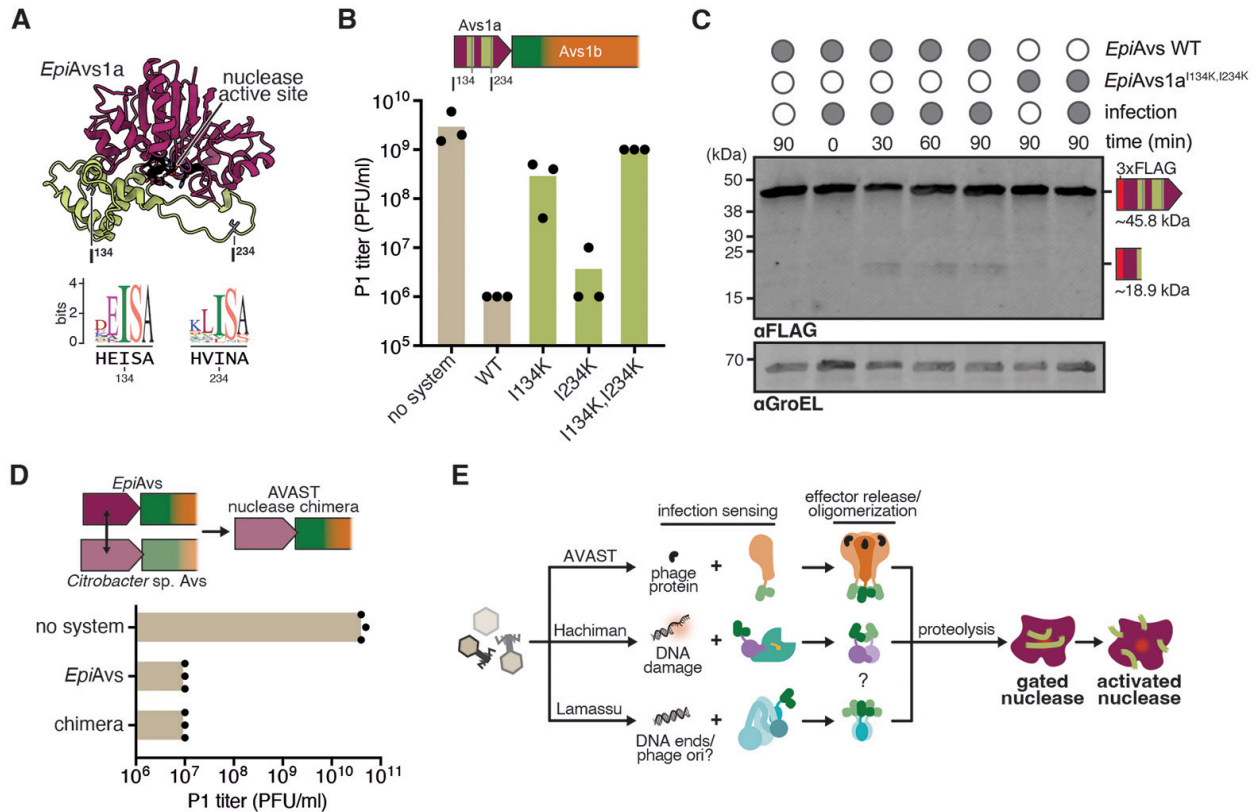


Figure 4-3. Nuclease-protease pairs function in AVAST. **A)** AlphaFold3 model of the AVAST Avs1a nuclease from *E. piriflorinigrans* (CFBP5888), with active site colored black, predicted protease cut sites colored gray, and insertions colored lime green. Below, sequence logos of predicted cut site positions among related nucleases, with the *EpiAvs1a* primary sequence as a reference. **B)** Quantification of phage P1 plaque assays against *EpiAvs* and predicted protease cut site mutants. Individual data points of three independent biological replications are shown along with the mean. A portion of the locus showing locations of insertions in the *EpiAvs1a* primary sequence and predicted cut sites is shown. **C)** Immunoblot of cell lysates from cells expressing *EpiAvs* or *EpiAvs* with cut site mutants in the Avs1a nuclease (I134K, I234K) infected with phage P1 (MOI 3). Both constructs encode a 3xFLAG-tag fused to the N-terminus of Avs1a. GroEL serves as a loading control. The full length protein (~45.8 kDa) and the predicted cleavage product (18.9 kDa) are indicated. **D)** Quantification of phage P1 plaque assays against *EpiAvs* and a chimeric *EpiAvs* system with an Avs1a nuclease from *Citrobacter* sp. ESBL3 (NZ_JAZEZV010000002.1). Individual data points of three independent biological replications are shown along with the mean. **E)** Model for a conserved mechanism of pro-nuclease activation among AVAST, Hachiman, and Lamassu defense systems. Upon infection sensing, defense system components oligomerize, triggering the protease and subsequent proteolytic removal of gating loops in co-encoded nucleases.

Avs1b activates a ‘gated’ effector nuclease via proteolysis, despite considerable phylogenetic distance from PflHamMAB and a different infection sensing mechanism (**Fig. 4-1B**). To demonstrate transferability of nuclease-protease systems implied by phylogenetic analyses, we constructed a chimeric *EpiAvs* encoding a nuclease

transplanted from a *Citrobacter* Avs system. This chimeric system displays equivalent defense against phage (**Fig. 4-3D**). Conservation and transferability of the nuclease activation mechanism implies that this gene pair behaves similarly in other antiphage systems (**Fig. 4-3E**). Consistent with this prediction, oligomerization upon infection sensing has been shown to activate fused effectors in Hachiman (**Fig. 4-2K**)²³⁷, AVAST⁴⁴ and Lamassu systems^{238,239} (**Fig. 4-3E**). Furthermore, disordered loops proximal to the active site are present in every antiphage trypsin-associated nuclease identified in this study (**Supplementary Fig. 4-19**).

4.3.5 Identification of a caspase-nuclease defense system

Gated nucleases co-occur with fused trypsin proteases in multiple known defense systems (**Fig. 4-1A,B**). Our phylogenetic analyses unexpectedly revealed an additional radiation of nuclease homologs that co-occur with proteases belonging to other diverse families (**Fig. 4-4A; Supplementary Fig. 4-2**). The largest clade in this group includes a nuclease-encoding gene positioned downstream of a gene encoding a protease most closely related to caspases, which are essential to human innate immunity and apoptosis (18) (**Fig. 4-4A,B**). These loci are variable in organization, encoding several genes of unknown function and multiple domains associated with oligomerization and protein-protein interactions, including tetratricopeptide repeats (TPR) and immunoglobulin-like (ig-like) folds (**Fig. 4-4B**)²⁴¹. Caspase-associated nucleases appear in immune hotspots known as ‘defense islands’ and near transposases, both signatures of immune function (**Supplementary Fig. 4-2C**)⁶⁶. We wondered whether these caspase-nuclease (*canu*) loci encode a similar mechanism, wherein the caspase activates its linked nuclease to provide antiphage defense.

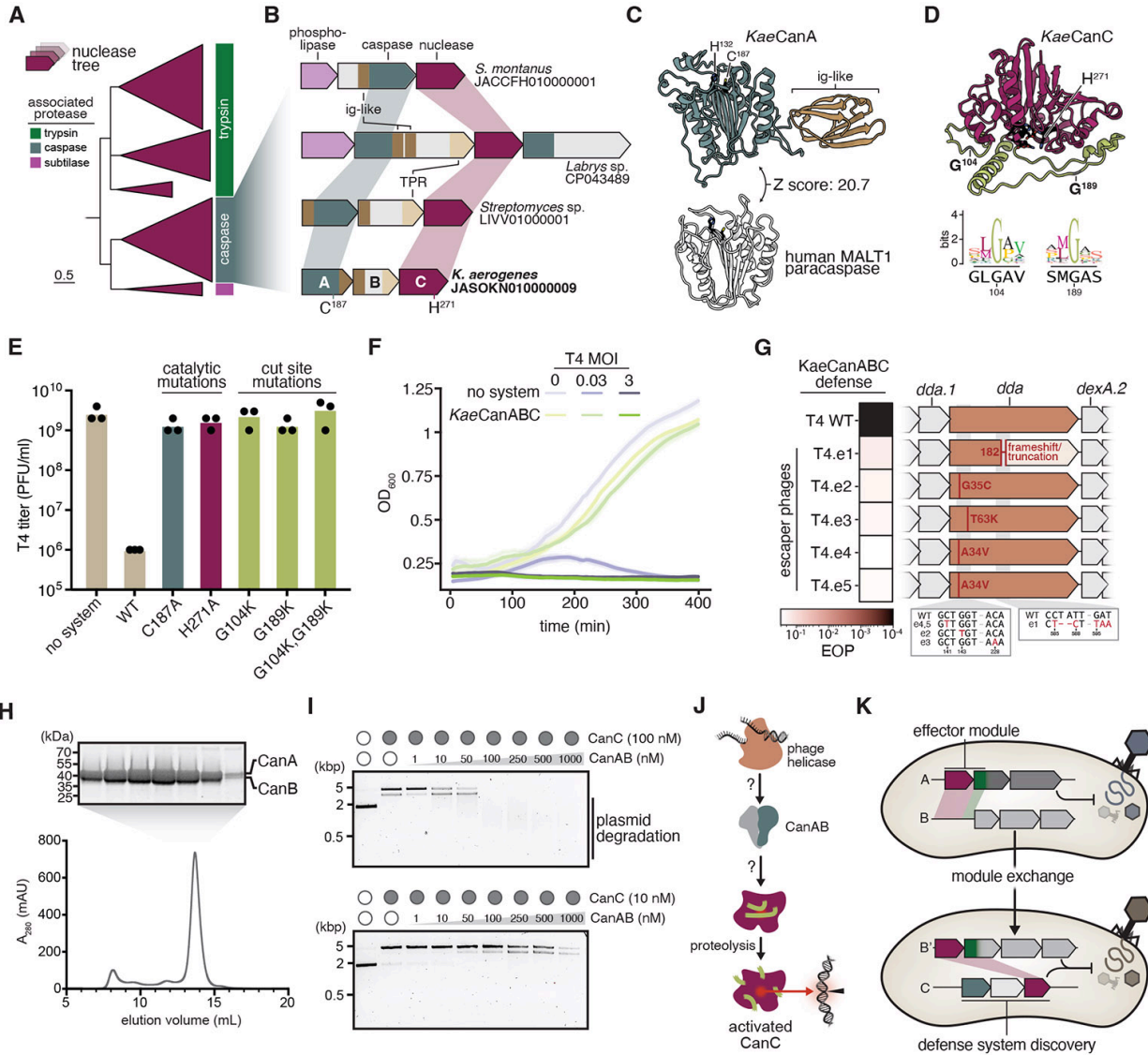


Figure 4-4. Canu is an antiphage defense system with a caspase-activated nuclease. **A)** Phylogenetic tree showing major clades of nucleases associated with multiple protease families, as annotated by the outer track (right). A subtree is shown. For the complete phylogeny, see **Supplementary Fig. 4-2**. **B)** Representative caspase-nuclease or *canu* loci with domains annotated (ig-like, immunoglobulin-like; TPR, tetratricopeptide repeat). The *Klebsiella aerogenes* CanABC operon (bottom) is annotated with predicted catalytic residues. **C)** Structural homology (DALI Z-score: 20.7) between the AlphaFold3 model of KaeCanA and the human paracaspase MALT1 (Mucosa-associated lymphoid tissue lymphoma translocation protein 1, PDB: 3UOA), with active sites colored black. **D)** AlphaFold3 model of KaeCanC, with active site colored black, predicted protease cut sites colored gray, and insertions colored lime green. Below, sequence logos of predicted cut site positions among related nucleases, with the KaeCanC primary sequence below. **E)** Quantification of phage T4 plaque assays against KaeCanABC and mutants shown in (A). Individual data points of three independent biological replicates are shown along with the mean. **F)** Growth curves of *E. coli* expressing KaeCanABC and an empty vector control during T4 infection at the specified MOI. Data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates. **G)** Characterization of phage T4 mutants that escape KaeCanABC defense. Left, heatmap showing efficiency of plaquing (scale below) of

wildtype phage T4 and five isolated escapee T4 mutants. Right, corresponding genotypes of phage T4 and escapers. Amino acid substitutions are shown on loci, and DNA-level detail is shown below. All escape mutations map to the *dda* gene. **H)** SDS-PAGE gel of fractions from indicated peak in gel filtration analysis of *KaeCanAB* coexpression. MBP-tagged CanA coelutes with CanB. **I)** Titration of CanAB with constant concentrations of *KaeCanC* (top: 100 nM, bottom: 10 nM) incubated with plasmid. **J)** Proposed model of *canu* defense. The CanAB complex may either directly or indirectly sense phage helicase. Upon activation, CanA cleaves the CanC pro-nuclease, leading to activation and DNA degradation. **K)** Model for the exchange of the nuclease-protease pair (top) and a strategy to use coevolution and transfer to discover new defense loci (bottom).

In the prototypical three-gene *canu* system from *Klebsiella aerogenes* (*KaeCanABC*), the *canA* gene encodes a caspase domain structurally similar to the human MALT1 paracaspase, retaining the C-terminal Ig-like fold crucial for immune activation (**Fig. 4-4B,C**)^{242,243}. The CanC nuclease active site is likely blocked by dual disordered insertions based on structure prediction and comparison to related enzymes (**Fig. 4-4D; Supplementary Fig. 4-18**). Both gating insertions contain conserved glycine residues (**Fig. 4-4C**), leading us to hypothesize that these positions are recognized by the caspase to trigger nuclease activity.

To test Canu activity, we challenged *E. coli* heterologously expressing *KaeCanABC*. We found that *KaeCanABC* provides strong protection against phage T4 (**Fig. 4-4E; Supplementary Fig. 4-20**). Single amino acid substitutions of predicted active site residues in the caspase or nuclease domains abolished defense, suggesting that both catalytic activities are required. Mutation of either predicted proteolysis site in the nuclease was also sufficient to ablate defense. The *KaeCanABC* system overcame infection at low viral doses, but succumbed to phage when viral titer surpassed that of the host (**Fig. 4-4F**). To identify potential viral triggers of Canu, we isolated mutant phages that escape *KaeCanABC* defense (**Fig. 4-4G; Supplementary Fig. 4-21**). Sequencing of escaper phages revealed diverse missense and frameshift mutations mapping to the *dda* gene, which encodes a nonessential SF1 helicase involved in phage DNA replication, recombination, and repair²⁴⁴. Together, these data imply that *KaeCanABC* carries a conserved pro-nuclease which, upon infection sensing involving phage helicase and proteolytic processing by its partner caspase, protects cells from infection by programmed cell death or growth arrest. Whether *KaeCanABC* recognizes phage helicase by direct recognition⁴⁴, or indirectly via their activities remains to be seen^{124,163}.

CanA is a caspase family protease and CanC is a putative pro-nuclease. To investigate Canu activities and determine a role for CanB, we purified each protein individually. Only CanC was obtained in isolation. Coexpression of the CanA and CanB produced a soluble complex, suggesting CanB plays a potential role in modulating CanA protease activity (**Fig. 4-4H**). Upon incubating CanC with increasing concentrations of CanAB, we observed conversion of CanC into a potent nuclease (**Fig. 4-4I**). Combined, these data demonstrate the caspase CanAB complex plays a role in activating the CanC nuclease for phage defense (**Fig. 4-4J**).

4.4 Discussion

In this study, we discovered that multiple different prokaryotic antiviral pathways employ related nuclease-protease enzyme pairs to diversify immune function. Cell-based and biochemical data showed that protease-catalyzed cleavage of a proenzyme form of the nuclease converts it to an active DNase. Once activated, the nuclease destroys cellular DNA, likely killing or arresting infected cells to halt viral replication. Unlike most cell death effectors that act individually^{43,46,58–60}, nuclease and protease genes function in tandem and co-evolve. This enzymatic partnership provides regulation that may safeguard against autoimmunity. For example, independent control over nuclease expression enables fine-tuning of its effect on cells, while fusion of partner proteases to core defense genes which require oligomerization to activate may prevent nonspecific activation^{60,196,238,239}. Our results reveal that proteolytic activation, long recognized as central to eukaryotic immunity^{245,246}, also shapes prokaryotic defense systems more extensively than previously appreciated^{247–249}.

The HamA effector convergently gained effector modules and lost canonical nuclease activity. One possible explanation for repetitive HamA active site residue depletion is that effector acquisition could lead to selection against HamA activity. Mutational drift in response to autoimmunity has been observed elsewhere in bacterial and animal immunity^{250,251} and is one possible explanation for the emergence of fusion-dHamA systems, but it remains unclear under which selective pressures module acquisition occurs. The genetic mechanism of exchange is equally mysterious.

Why do some defense systems appear to interact and exchange modules more than others? We propose that the arrangement of genes within systems is a contributing factor. The Lamassu system was recently shown to respond to phage DNA by liberating the upstream LmuA effector, which then multimerizes to trigger activity^{238,239}. AVAST recognizes specific phage proteins and multimerizes, activating a variety of N-terminal effectors^{44,252}. In Hachiman, the HamA effector is also thought to oligomerize upon HamB activation¹²⁴. In all three systems, infection sensing induces multimerization of an effector encoded upstream. Furthermore, the different domains fused to HamA, LmuA and Avs themselves multimerize as an obligate step in activation, as evidenced by cJun fusion experiments here and elsewhere. This is also the case for Cap4^{60,196} and TIR domains^{253–255}. A shared logic of effector activation, combined with a similar operon arrangement, may enable facile acquisition of new effectors with minimal impact on system fitness (**Fig. 4F**).

Our phylogenetic analyses suggest that ancestral nucleases acquired autoinhibitory insertions containing endopeptidase sites, creating a requirement for protease partners to restore activity. The resulting protease-nuclease pairs were then transferred between otherwise unrelated antiviral systems. Despite frequent mobilization, these enzyme pairs retain the coupled mechanism of nuclease activation, reflecting their common ancestry. Distinct genetic arrangements, including relative position and standalone

versus fused configurations in trypsin and caspase systems, respectively, likely trace back to this fundamental functional association. Based on these observations, it is possible that other proteolytically activated enzyme classes participate in antiviral immunity, providing similar regulatory safeguards. Hallmarks of functional coupling, including protease coevolution and active site-blocking insertions, can guide the discovery of additional enzymatic partnerships (**Fig. 4-4K**).

In this work, we demonstrate that new immune loci may be found by tracing modular effector genes throughout prokaryotic genomes, presenting caspase-nuclease (canu) antiphage loci. Canu systems which extend the diversity of known caspase-containing immune functions in prokaryotes^{241,256–258}. In humans, multiple caspase paralogs orchestrate programmed cell death by either proteolytically activating other factors (initiator caspases), or by directly targeting cellular proteins (inflammatory and executioner caspases)^{245,259,260}. Prokaryotes encode other caspase homologs with known functions in antiphage defense^{256,258}, including Type III CRISPR^{247–249}, gasdermin/CARD^{261,262}, NLR⁴⁵ and Thoeris systems²⁶³. Canu systems differ by encoding Ig-like domains that also occur in human caspases^{241,242}. The canu system from *K. aerogenes* (KaeCanABC) provides potent phage defense and shows striking conservation of the nuclease activation mechanism found in Hachiman and AVAST. These results suggest that the Canu caspase functions as a specific endopeptidase, analogous to eukaryotic initiator caspases²⁶⁴. While canu diversity and functions await further exploration, conservation of nuclease activation provides a basis for determining the mechanism of canu defense.

4.5 Methods

4.5.1 Nuclease phylogenetic and coevolutionary analyses

Phylogenetic analyses of MBL hydrolase-like nucleases were initiated by generating a seed alignment of HamM, Avs1a and LmuA genes using ClustalOmega²⁰⁶. The resulting alignment was manually trimmed and then used to generate a custom HMM with hmmbuild²⁶⁵, and subsequently hmmsearch was performed on the uniprot database (2025_01). Unique HMM hits with an e-value less than 1 E-3 were selected, and clustered with MMseqs2 easy-linclust using --min-seq-id 0.9²¹⁰. Multiple sequence alignment using ClustalOmega was performed, followed by manual trimming. Finally, IQ-TREE was performed with the following parameters: -bb 1000 -nm 1500 -safe, with ModelFinder used to select VT+R10^{266,267}. The phylogenetic tree was visualized with iTOL²⁶⁸, with branches with bootstrap values below 60 deleted. A large clade containing the ComEC hydrolase/nuclease domain was chosen as a root. A subtree of only trypsin-like protease clades is shown in **Fig. 4-1A**, and a cartoon subtree of all protease-associated domains is shown in **Fig. 4-4A**.

Context analysis was performed by extracting 10 kbp up and downstream of hmmsearch hits using the NCBI Entrez tool²⁶⁹. PADLOC was then used to annotate defense systems within the gene neighborhood¹⁵³. If hit proteins were encoded within a known defense system as determined by PADLOC, these were annotated as such in phylogenetic trees using iTOL annotation tools²⁶⁸. Leaves without defense system annotations were manually inspected and labelled. Nuclease structures from *Pfl*HamM, *Epi*Avs1a, *Bce*LmuA-H, and UG9 DRT systems were predicted using the AlphaFold3 web server²³⁶. Structures were aligned and visualized with matchmaker in ChimeraX²⁷⁰.

Nuclease-protease coevolutionary analysis was performed using a filtered set of nuclease-proteases with genomic contexts of at least 20 kbp (n=191). Protein coding genes found within the neighborhood were pooled, and clustered using MMseqs easy-cluster on default parameters. To further group clusters, HHblits was used on default parameters²⁷¹. Clusters were merged when both query-and-target coverage were above 50, and probability was above 90. Most trypsin proteases (n=164) fell under a single group. Remaining proteases that fell outside the cluster (n=36) were retrieved by an hmmsearch using an alignment of the clustered proteases generated using MAFFT- E-INS-I²⁷². One nuclease containing locus did not have a clear trypsin protease homolog, and was excluded from further analysis. Alignments of nucleases and proteases were made using MAFFT-E-INS-I, and nuclease-protease pairs were removed from further analysis if either genes appeared to be incomplete or inactive. This resulted in a set of 181 nuclease-protease pairs used as input for nuclease and protease trees. To generate a species tree, the latest taxonomic metadata and phylogenetic tree (bac120.tree) were retrieved from the GTDB database. The bac120.tree was further pruned to only display leaves associated with relevant genomes containing nuclease-protease pairs (n=160; 21 pairs had no match on GTDB).

To qualitatively assess coevolution, the phylogenetic topology of the nuclease and protease were compared to each other and to the species tree. First, the nuclease and protease alignments were trimmed and passed to IQ-TREE to generate trees for each of the respective genes using: -m MFP -bb 1000 -alrt 1000. Next, the resulting gene trees and the pruned species tree were passed to the program CoPhylo in pairs for comparison²⁷³. This was done for the following pairs: nuclease vs. protease, nuclease vs. species, and protease vs. species. We then generated tanglegram visualizations in iTOL. Further reconciliation analysis was conducted for the gene trees versus the species tree. The nuclease and protease alignments were concatenated into a single alignment file. This concatenated alignment file and the species tree were passed to the program GeneRax to compute reconciliation likelihood, phylogenetic likelihood, and joint likelihood. The following parameters were selected: -r UndatedDTL and -geneSearchStrategy SPR and LG+G substitution model²⁷⁴.

4.5.2 Hachiman HamA phylogenetic analysis

For analysis of Hachiman diversity, unique proteins from InterPro protein family IPR014976 and annotated as Hachiman antiphage defense system protein HamA in Refseq and Genbank were pooled and chosen for phylogenetic analysis. MMseqs2 easy-linclust was run on pooled sequences with --min-seq-id 0.552²¹⁰. Then, an HMM was built based on the clustered sequences with hmmbuild, and subsequently hmmsearch was performed on the genbank nr nucleotide database (v274)²⁶⁵. Unique HMM hits with an e-value less than 1E-5 were selected, and clustered with MMseqs2 easy-linclust using --min-seq-id 0.9. Multiple sequence alignment using muscle5.1 -super5 was performed on the clustered sequences, with an additional 10 homologs of the *P. aquatilis* type IIS restriction endonuclease domain chosen as the outgroup^{124,275}. Truncated Hachiman variants (HamAs that are <150 amino acids) were manually removed and HamA-HamB fusions (likely due to stop codon misannotation) were removed from the sequence alignment, then re-aligned with muscle5.1 -super5. Poorly aligned regions were trimmed by trimAl with -gt 0.26²⁷⁶. Finally, IQ-TREE was performed with --bb 1000 -m LG+F+G4 (50). The phylogenetic tree was visualized with iTOL²⁶⁸, with branches with bootstrap values below 60 deleted and branches annotated for taxonomy and the presence or absence of a fused domain (which were trimmed from the alignment to observe only HamA domain dynamics).

HamA were plotted by length, then uncommonly long HamAs were extracted and aligned. Representatives of each long HamA sequence were folded using Colabfold²⁰³, then fused domains were manually extracted in ChimeraX²⁷⁰. Each of the five extracted N-terminal domains were used to search for structural homologs via the Dali search server²⁷⁷. Superimpositions were generated using the matchmaker function in ChimeraX²⁷⁰. Sequence logos and locus visualizations were generated using Geneious Prime. Phylogenetic analyses of occurrences of HB and YTH domains outside of Hachiman systems were executed in the same manner as HamM phylogenetic analysis described above.

4.5.3 Bacterial strains and phages

For bacterial cultivation, *E. coli* strains maintained as 25% (v/v) glycerol stocks and stored at -80°C were generally grown in LB Lennox media at 37°C and 250 rpm aeration. Media was supplemented with carbenicillin (Sigma, 100 µg/mL) or chloramphenicol (Sigma, 20 µg/mL) where applicable. *E. coli* Top10 (Invitrogen) was used as a cloning strain. For phage experiments using the PflHamMAB system, *E. coli* MDS42 was used. For *EpiAvs* and *KaeCanABC* phage experiments, we used *E. coli* K-12 MG1655. To purify proteins, *E. coli* BL21-AI (Invitrogen) was used.

Phage P1 was obtained from the American Type Culture Collection (ATCC 25404-B1). Phages from the BASEL collection were kindly provided by Dr. Alex Harms²⁷⁸. All phages were verified through either whole-genome sequencing or targeted genotyping

to check for stock cross-contamination. Wildtype phages were propagated at 37°C in LB Lennox media using an initial MOI of approximately 0.1 and host *E. coli* BW25113. After culture collapse, a drop of chloroform was added and cultures were centrifuged at 4000 g for 15 min. The supernatant was passed through a 0.2 µm filter and stored at 4°C. All phage titers were determined on assay hosts with an empty vector control in the p15a backbone using plaque assays (described below).

4.5.4 Plasmid construction

All plasmids used here were constructed using Gibson Cloning, Golden Gate Assembly or mutagenic PCR/ ligation, using PCR products purified by gel extraction (Zymo D4001). *Pseudomonas fluorescens*, *Klebsiella aerogenes*, and *Citrobacter* sp. ESBL3 loci were synthesized by Twist. For phage assays, all defense loci except AVAST were cloned under control of the pTet promoter in a P15A backbone encoding for chloramphenicol resistance. In addition to the pTet promoter, the *KaeCanABC* is also under the control of its native promoter, and experiments with this system were performed with no additional induction of protein expression. The AVAST system, cloned from the source plasmid⁴³, is under control of the *bla* promoter in a p15a backbone, which encodes chloramphenicol resistance. For phage terminase coexpression, the pJEX vector conferring kanamycin resistance was used for inducible control of genes upon addition of crystal violet. For protein purification purposes, loci were cloned under a T7 lactose-controlled promoter in a high copy vector encoding carbenicillin resistance. To assess the toxicity of HamM activated by dual-loop proteolysis (henceforth ‘pre-cut HamM’), we pre-programmed cuts in HamM by inserting stop codons immediately after the putative cut-site loop isoleucine residues. Concurrently, we added strong ribosome binding sites (BBa_B0034) and a start codon before resumption of the coding sequence²⁷⁹. The resulting tripeptide is predicted by AlphaFold3 to retain the nuclease tertiary structure. For transplantation experiments, a nuclease from a *Citrobacter* sp. ESBL3 (WP_332628315.1) was incorporated into the *EpiAvs* locus by Gibson Assembly.

4.5.5 Plaque assays

Plaque assays were performed as previously described, using the double agar overlay method¹²⁴. In brief, 100 µL (AVAST, *KaeCanABC*) or 200 µL (*PflHamMAB*) of saturated overnight cultures carrying either an empty vector or the defense system of interest were added to molten LB Lennox agar (0.7% w/v agar, 60°C – henceforth, “top agar”), then the mixture was supplemented with antibiotics. For plaque assays assaying the antiviral activity of *PflHamMAB* against the diverse panel of phages, top agar was supplemented with anhydrotetracycline (aTc, Sigma) at a final concentration of 0.5 nM to induce expression of HamMAB. For plaque assays assaying the antiviral activity of *PflHamMAB* mutants against EdH4, top agar was supplemented with anhydrotetracycline (aTc, Sigma) at a final concentration of 1 nM to induce expression of HamMAB. For plaque assays assaying the antiviral activity of the AVAST and

KaeCanABC systems, no additional inducers were added to the top agar. This mixture was poured over 1.5% w/v LB Lennox agar plates (presupplemented with antibiotic) and was allowed to cool and solidify. Phages were diluted tenfold in SM buffer (Teknova), then 2 or 2.5 μ L of eight dilutions were spotted onto the top agar layer. For phage P1, a stock approximately tenfold higher in titer than that used for **Fig. 4-3B** is shown in **Fig. 4-3D**. After drying under sterile conditions, plates were incubated at 30°C for 12-16 hours. Plates were then scanned and plaque forming units (PFU) were enumerated. Plaque assays were performed in biological triplicate, and data were analyzed using Graphpad Prism and Python.

4.5.6 Liquid infection assays

To assay bacterial growth in the presence or absence of phage, OD600 was monitored in a 96-well plate format in a plate reader (Biotek Cytation 5). To initiate growth, saturated overnight cultures carrying either an empty vector or the defense system of interest were seeded in a microplate (Corning 3903) at a CFU of $\sim 8 \times 10^6$ CFU per well in 200 μ L of LB media supplemented with antibiotic and, when assaying the antiviral activity of PflHamMAB, with aTc at a final concentration of 0.5 nM or 0.125 nM. Phage were added to achieve the indicated MOI after separate dilution in SM buffer. Growth was monitored by measurement of OD600 every 5 minutes while shaking at 800 RPM (double orbital) at 30°C for the indicated time span. Liquid growth assays were performed in biological triplicate, and data were analyzed using Python.

4.5.7 Protein expression and purification

For complex purification vectors in the native locus format, only HamA and CanA were tagged with 10xHis-MBP-TEV. After transformation into BL21-AI *E. coli*, single colonies were inoculated in TB media and were grown to an optical density of ~ 0.6 then induced overnight at 16°C with 0.5 mM isopropyl- β -D-thiogalactopyranoside (IPTG) and 0.1% L-arabinose. Cells were harvested and resuspended in lysis buffer (20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.7, 500 mM NaCl, 20 mM imidazole, 0.1% Triton X-100, 1 mM Tris (2-carboxyethyl)phosphine (TCEP), Complete EDTA (ethylenediaminetetraacetic acid)-free protease inhibitor (Roche), 0.5 mM phenylmethylsulfonyl fluoride (PMSF) and 10% glycerol). Cells were lysed by sonication, then clarified by centrifugation. The clarified lysate was incubated with Ni-NTA (Qiagen) resin for 1 hr. The resin was washed with wash buffer (20 mM HEPES, pH 8, 150 mM NaCl, 20 mM imidazole, 1 mM TCEP, and 5% glycerol), then bound protein was eluted with wash buffer supplemented with 300 mM imidazole. Eluate was then run over an MBPTrap column (GE Healthcare), washed with MBP/SEC wash buffer (20 mM HEPES, pH 8, 150 mM KCl, 1 mM TCEP, and 5% glycerol), and eluted with MBP/SEC buffer supplemented with 10 mM maltose. Eluted protein from the MBPTrap column was treated with TEV protease overnight. Then, TEV protease-treated samples were run through Ni-NTA resin, and flow-through was collected. Flow-through was

concentrated and run on a Superdex 200 10/300 GL column (Cytiva). Aliquots were concentrated and snap-frozen in liquid nitrogen for later use.

4.5.8 Protease-HamAB proteolytic cleavage assays

Protease cleavage assays of HamM by protease-HamAB were performed in isothermal amplification buffer (20 mM Tris-HCl, pH 8.8, 10 mM (NH)₂SO₄, 50 mM KCl, 2 mM MgSO₄, 0.1% Tween 20). 1.4 μM of protease-HamAB was incubated with 1 μM of HamM for 90 minutes at 37°C. Controls were performed with protease-HamAB only and HamM only as well. Reactions were terminated with addition of 4x Laemmli buffer. Samples were boiled for 3 minutes at 95°C, loaded on a 4-20% SDS PAGE gel, stained by Coomassie InstantBlue, and imaged on a ChemiDoc MP (BioRad). To assay effects of ATP and activating DNA on proteolytic cleavage of HamM, the assay above was modified as follows. 0.5 μM of protease-HamAB was mixed with ATP at the indicated concentration or 5 mM AMP-PNP (Sigma, Adenylyl-imidodiphosphate). Activating dsDNA with 3' overhang was added to a final concentration of 2 μM where indicated. Reactions were allowed to incubate for 10 minutes at 30°C before introduction of HamM. After addition of HamM, the reaction was allowed to proceed for 30 min, after which it was quenched and analyzed as described above.

Protease cleavage assays of SSB (Fisher Scientific), RecBCD (New England Biolabs), and Topoisomerase I (New England Biolabs) by protease-HamAB were performed in isothermal amplification buffer. All reactions were performed with a final concentration of 0.25 μg/μL SSB, 0.5 U/μL RecBCD, and 0.25 units/μL Topoisomerase I. For the indicated reactions (**Supplementary Fig. 4-8**), SSB, RecBCD, and Topoisomerase I were boiled at 95°C for 3 minutes before addition of protease-HamAB. Reactions were started by the addition of protease-HamAB to a final concentration of 0.5 μM and incubated at 30°C for 30 minutes. Controls with SSB, RecBCD, and Topoisomerase I only were conducted as well. Reactions were stopped by addition of 4x Laemmli buffer. Samples were boiled for 3 minutes at 95°C, loaded on a 4-20% SDS PAGE gel, stained by Coomassie InstantBlue, and imaged on a ChemiDoc MP (BioRad).

4.5.9 In-gel mass spectrometry

Experiments for mass spectrometry were performed in MBP/SEC buffer supplemented with 1 mM MgCl₂. 9.6 μM of protease-HamAB was incubated with 19.3 μM of HamM for 60 minutes at 30°C. Reactions were terminated with addition of 4x Laemmli buffer. Samples were heated for 3 minutes, loaded on a 4-20% SDS PAGE gel, and visualized by Coomassie InstantBlue. Protein bands corresponding to proteolytic products of HamM were excised with a razor into 1 mm² cubes. Gel pieces were washed twice with 50% (v/v) acetonitrile and 50 mM ammonium bicarbonate (pH 8) for 15 minutes with shaking. Gel pieces were then dehydrated with 100% acetonitrile for 5 minutes with shaking. Then the solvent was removed, and gel pieces were allowed to air dry for 20 min. To the dry gel pieces, 10 mM TCEP and 40 mM chloroacetic acid were added and

incubated at 70°C for 5 min. The gel pieces were washed again with 50% acetonitrile and 50% 50 mM ammonium bicarbonate for 15 min with shaking. The gel pieces were then rehydrated in 50 mM ammonium bicarbonate and 1 µg chymotrypsin/ GluC (1:50) was added and incubated for 1 hour at room temperature, then 50 mM HEPES pH 8 solution was added to cover the pieces. The samples were allowed to incubate overnight at 37°C. Peptides were extracted from gel pieces with 25% acetonitrile and 50 mM ammonium bicarbonate pH 8, then again with 100% acetonitrile for 5 min with shaking. Samples were filtered through a 0.22 µm PVDFspin column (Millipore) Peptides were dried in a speedvac apparatus to 30-60 µL total volume, then samples were acidified with 2 µL formic acid (neat).

In-gel chymotrypsin/GluC digested peptides were analyzed by online capillary nano LC-MS/MS using a 25 cm reverse phase column and a 10 cm precolumn fabricated in-house (75 µm inner diameter, packed with ReproSil-Gold C18-1.9 µm resin (Dr. Maisch GmbH)) that was equipped with a laser-pulled nanoelectrospray emitter tip. The precolumn used 3.0 µm packing (Dr. Maisch GmbH)). Peptides were eluted at a flow rate of 300 nL/min using a linear gradient of 2–40% buffer B in 140 min (buffer A: 0.05% formic acid and 5% acetonitrile in water; buffer B: 0.05% formic acid and 95% acetonitrile in water) in an Thermo Fisher Easy-nLC1200 nanoLC system. Peptides were ionized using a FLEX ion source (Thermo Fisher) using electrospray ionization into a Fusion Lumos Tribrid Orbitrap Mass Spectrometer (Thermo Fisher Scientific). Instrument method parameters were as follows: MS1 resolution, 120,000 at 200 m/z; scan range, 350–1600 m/z, orbitrap mode acquisition. The top 20 most-abundant ions were subjected to collision-induced dissociation with a normalized collision energy of 35%, activation q 0.25, and precursor isolation width 2 m/z. Dynamic exclusion was enabled with a repeat count of 1, a repeat duration of 30 seconds, and an exclusion duration of 20 seconds. RAW files were analyzed using PEAKS (Bioinformatics Solution Inc) with the following parameters: semi-specific cleavage specificity at the C-terminal site of R and K, allowing for 5 missed cleavages, precursor mass tolerance of 15 ppm, and fragment ion mass tolerance of 0.5 Daltons. Methionine oxidation was set as variable modifications and cysteine carbamidomethylation was set as a fixed modification. Peptide hits were filtered using a 5% FDR. Proteins with at least 2 unique peptides were filtered with a 5% FDR. Label free quantitation (LFQ) was performed using the PEAKS quantitation module and default parameters with the following exceptions: Top 2 peptides for each protein with a min of 10x10⁴ abundance was used and the TIC was used for all normalization including technical replicates. Peptide mass spectrometry data were visualized using PrIntMap-R²⁸⁰.

4.5.10 Plasmid and phage gDNA clearance assays

Plasmid and phage DNA interference assays were conducted in isothermal amplification buffer. The plasmid substrate is a generic pUC19 vector, and the phage DNA substrate is EdH4 genomic DNA isolated as described above. DNA was diluted to

4 ng/μl. Reactions were started with addition of Protease-HamAB at 5 nM and HamM at 10 nM, and incubated at 30°C for 1 min, 5 min, or 30 min. Reactions were quenched with addition of proteinaseK (New England Biolabs) and 10x Bluejuice. Reactions were imaged on 1% TAE agarose gels. Gels were stained with SYBR Safe and imaged on a ChemiDoc MP (BioRad).

For *KaeCanABC* plasmid cleavage experiments, CanC was diluted to the indicated concentrations in isothermal amplification buffer with plasmid DNA diluted to 4 ng/μl. CanAB was added at the indicated concentrations, and reactions were allowed to proceed at 37°C for 30 min before quenching by addition of proteinaseK (New England Biolabs) and 10x Bluejuice. Reactions were imaged on 1% TAE agarose gels. Gels were stained with SYBR Safe and imaged on a ChemiDoc MP (BioRad).

4.5.11 Protease-HamAB nuclease short nucleic acid activity assay

Short nucleic acid (ds/ssDNA, ds/ssRNA) cleavage assays were performed in isothermal amplification buffer or MBP/SEC buffer supplemented with 1 mM ZnCl₂. 0.5 μM of protease-HamAB and 0.5 μM of HamM were incubated with 0.5 μM nucleic acid for 30 minutes at 30°C. Reactions were stopped by addition of ProteinaseK (New England Biolabs) and 10x Bluejuice. Reactions were run on a 2% agarose gel stained with SYBR Gold (Thermo Fisher) and imaged on a ChemiDoc MP (BioRad).

4.5.12 Malachite green ATPase assays

Orthophosphate liberation was determined with a Malachite Green Phosphate Assay kit (BioAssay Systems) according to the manufacturer protocol. All reactions were conducted in isothermal amplification buffer. Substrates tested included two different ssDNAs (ssDNA 1 and ssDNA 2), blunt-ended dsDNA, dsDNA with 3' overhang, dsDNA with 5' overhang, and plasmid DNA. DNA substrates were diluted to 100 nM, or 4 ng/μL for plasmid reactions, in a total reaction volume of 40 μL in clear bottom, flat, black 96-well assay plates (Corning Costar). Reactions were started with addition of *Pfl*HamAB to 40 nM and ATP to 1 mM, followed by incubation at 30°C for 30 minutes. Additional controls with buffer only, ATP only, HamAB only, and no DNA substrates were performed. Reactions were quenched with the addition of activated malachite green reagent. The absorbance values of wells were measured after 20 min of color development at ambient temperature with a Biotek plate reader at 620 nm. Orthophosphate liberation was interpolated against a standard curve with known concentrations of free phosphate.

4.5.13 Nuclease and caspase structural analyses

Nuclease structures from *Pfl*HamM, pre-cut *Pfl*HamM, *EpiAvs1a* and *KaeCanC* were predicted using the AlphaFold3 web server²³⁶. Predicted structures were compared to the AlphaFold Database (AFDB) using Foldseek⁵¹. A confident hit, (uniprot: P39695) protein 3 from the *B. subtilis* ComE operon, was aligned and visualized with

matchmaker in ChimeraX²⁷⁰. Sequence logos of insertion regions containing the putative protease cleavage sites as determined by mass spectrometry (see above) were generated by performing BLAST searches, then aligning 100 hit sequences with ClustalOmega²⁰⁶ and visualizing regions of interest with the built-in sequence logo function of Geneious Prime. To generate a structural comparison of KaeCanuA, Dali was used to search the protein data bank for remote homology to experimentally determine structures, then superpositions were regenerated and visualized in ChimeraX using matchmaker.

4.5.14 Liquid toxicity assays

Liquid culture toxicity assays were performed in 96-well plate format in a plate reader (Biotek Cytation 5), and toxicity was assessed by monitoring OD600 over time. Saturated overnight cultures of the indicated strains were seeded into a microplate (Corning 3903) at a CFU of ~8e6 CFU per well in 200 μ L of LB media supplemented with antibiotics and the indicated concentration of inducer. For drug activation experiments, nalidixic acid (Sigma) or gentamicin (Sigma) were also added at the indicated sub-inhibitory concentrations as determined previously¹²⁴. Growth was monitored by measurement of OD600 every 5 minutes while shaking at 800 rpm (double orbital) at 30°C for the indicated time span. Liquid growth assays were performed in biological triplicate, and data were analyzed using Python.

4.5.15 Western blot analyses

To assess nuclease cleavage in cells, overnight cultures of either *Epi*Avs or the cut site mutant, each with an N-terminal 3xFLAG tag, were diluted 100-fold, then incubated at 30°C with shaking at 250 rpm until cultures reached an optical density of 0.5. Phage was then added at an MOI of 3. 300 μ L aliquots were taken at the indicated time points and were lysed with the addition of 1x Laemmli buffer with 50 mM TCEP. Samples were boiled for 3 minutes at 95°C and were then loaded on a 4-20% SDS PAGE gel. Separated proteins were transferred to a PVDF membrane (Bio-Rad, #1620177) for 1 hour at 100 V in Towbin buffer (25 mM Tris base, 192 mM glycine, 20% methanol). Membranes were blocked (5% (w/v) non-fat milk in 1x Dulbecco's Phosphate Buffered Saline with Tween-20 (DPBST)) for 1 hour at room temperature. After three washes in DPBST, membranes were incubated in primary antibody (1:1000 Mouse Anti-FLAG primary antibodies (Sigma, Catalog #F1804), 1:80,000 Rabbit Anti-GroEL primary antibody (Sigma, Catalog #G6532)) overnight at 4°C. After three washes in DPBST, membranes were then incubated in 1:20,000 secondary antibodies (Goat Anti-Mouse fluorescent secondary antibody 680RD (Licor, Catalog #926-68070), Goat Anti-Rabbit fluorescent secondary antibody 800CW (Licor, Catalog #926-32211) for 60 min at room temperature. After three washes with TBST, results were imaged on a LiCor at 700 nm and 800 nm emission wavelengths.

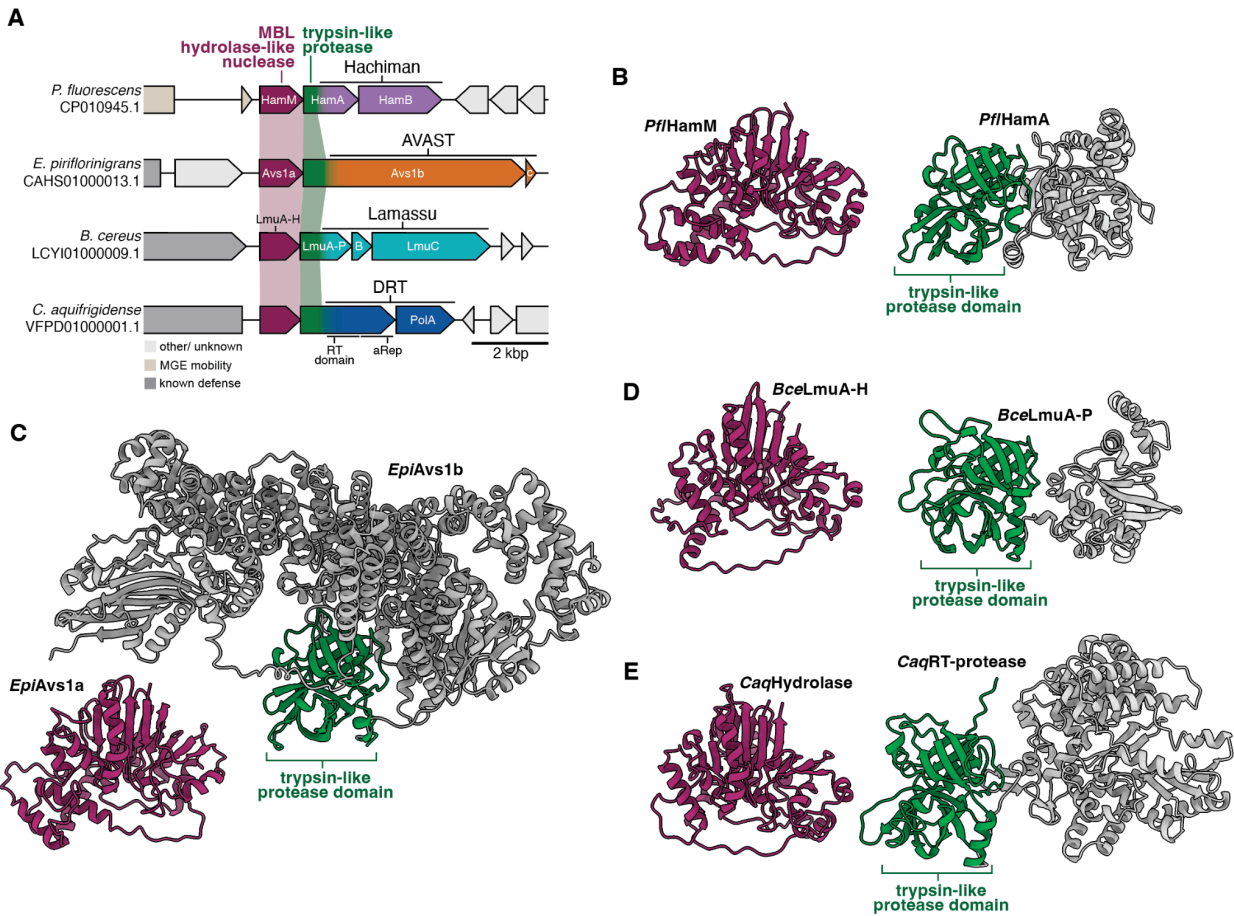
4.5.16 Isolation, scaling validation and sequencing of mutant phages that overcome defense

To isolate mutant phages that escaped *KaeCanABC*-mediated defense, plaque assays were performed as described above. After overnight incubation at 30°C, single plaques from the highest phage dilution where plaques were visible were picked and resuspended in 50 µL of SM buffer. To confirm the ability of these “escaper” phages to avoid targeting by the *KaeCanABC* system, putative escaper phages and the ancestor phage (i.e., a plaque formed by T4 on a lawn of cells encoding the empty vector) were further diluted in tenfold serial dilutions in SM buffer, and re-plaques on lawns of *E. coli* encoding either an empty vector or the *KaeCanABC* defense operon. After overnight incubation at 30°C, plaque forming units (PFUs) were counted and used to estimate the fold-defense (ratio of PFUs on empty vector to PFUs on *KaeCanABC*) for each phage.

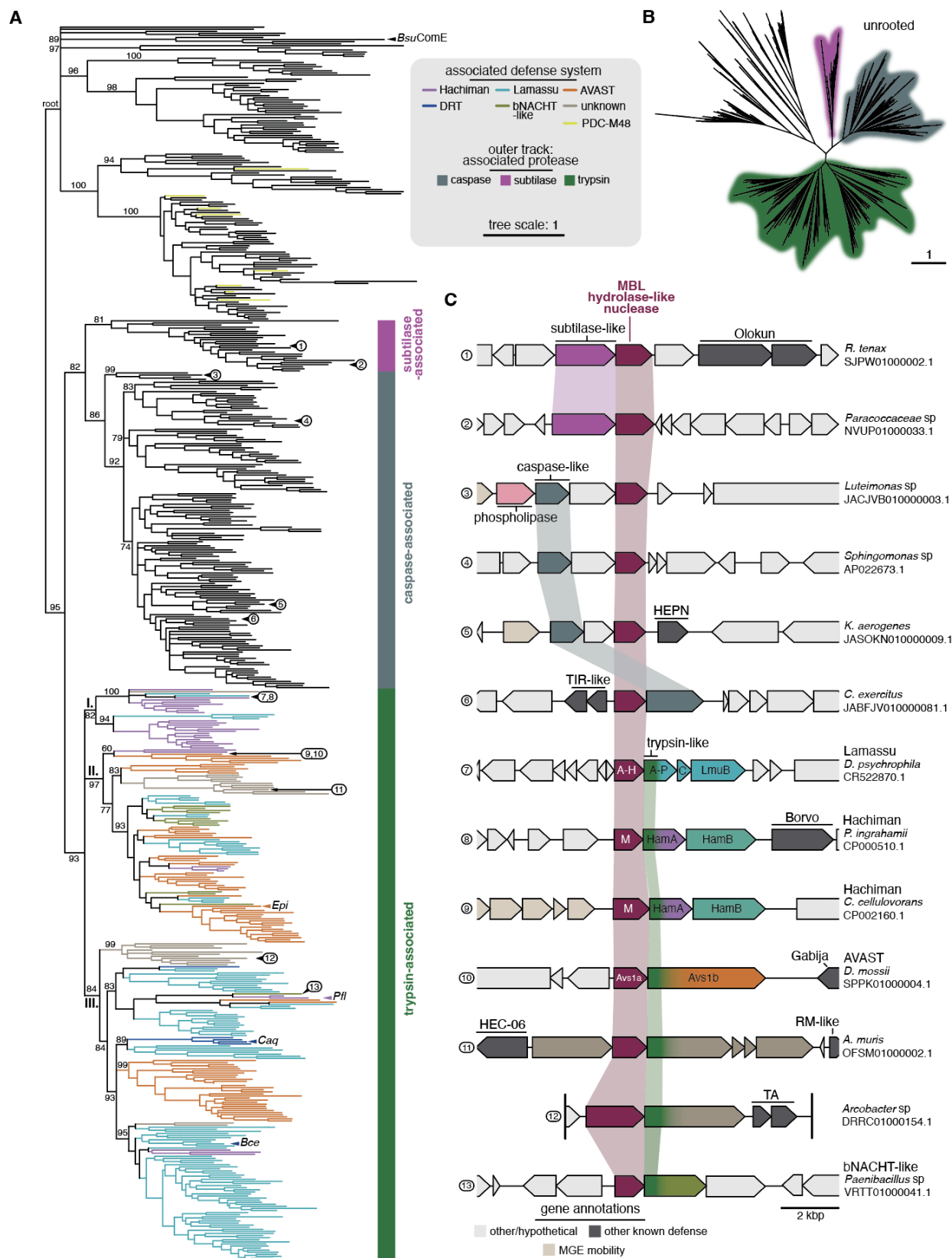
Escaper phages that plaqued similarly on cells expressing *KaeCanABC* and cells carrying the empty vector were further propagated to obtain high titre phage stocks for sequencing as described above. Phage genomic DNA was extracted from 200 µL of concentrated phage lysate using the Norgen Biotek Phage DNA Isolation Kit (# 46800), following the manufacturer’s instructions with the following minor modifications. Briefly, bacterial DNA was removed from the phage lysate by incubating it with 20 µL of RQ1 DNase I (PROMEGA M610A, 1 U/µL) at room temperature for 15 minutes, followed by a 5-minute incubation at 75°C to inactivate the enzyme. Then, the phages were lysed by treatment with 500 µL of Norgen Lysis Buffer supplemented with 10 µL of Proteinase K (New England Biolabs P8107S, 800 U/mL), and subsequent incubation at 55°C for at least 30 minutes, followed by inactivation of the enzyme by incubation at 65°C for 15 minutes. After addition of 320 µL of isopropanol, the lysate was bound to the Norgen column and gDNA purification followed as recommended by the manufacturer. gDNA was double-eluted from the column in a total of 300 µL of ddH₂O.

gDNA short read sequencing was performed by SeqCoast Genomics (short reads; 200 Mbp/1.3 million reads per sample). The ancestor phage reference genome was assembled by aligning reads to the phage T4 reference genome (GenBank: AF158101.6) to obtain a consensus sequence. Escaper phage genomes were assembled using this ancestor phage reference genome, and then re-mapped to it, allowing for the identification of mutations unique to escaper phages.

4.6 Supplementary Information

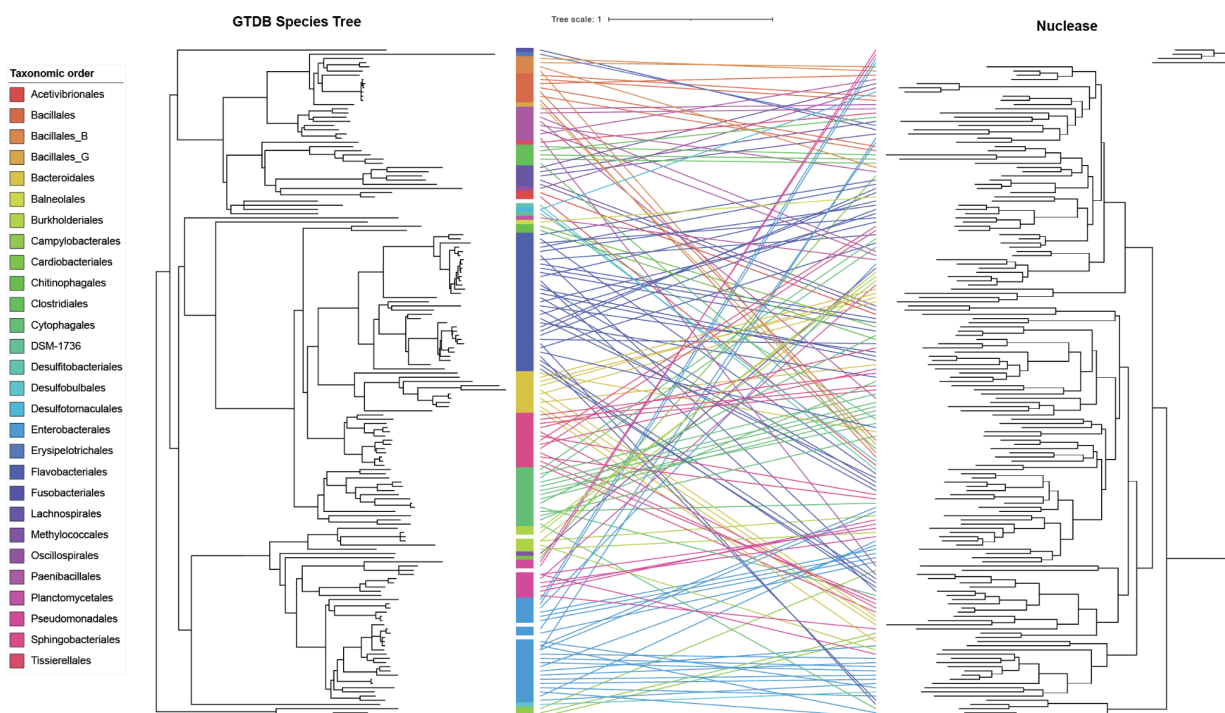


Supplementary Figure 4-1. Structural comparison of nuclease-protease pairs in known defense systems. **A)** Loci illustrating the genetic architecture of multiple defense systems encoding an N-terminal MBL hydrolase-like nuclease and partnering trypsin-like protease, as in **Fig. 4-1A**. Annotations for gene context and gene names are shown. **B-E)** AlphaFold3 predicted protein structures for nuclease and protease-fusion genes in Hachiman, AVAST, Lamassu, and DRT (UG9) systems. In each case, protein structures were predicted using loci from (A). Predicted structural snapshots are aligned to show homology.



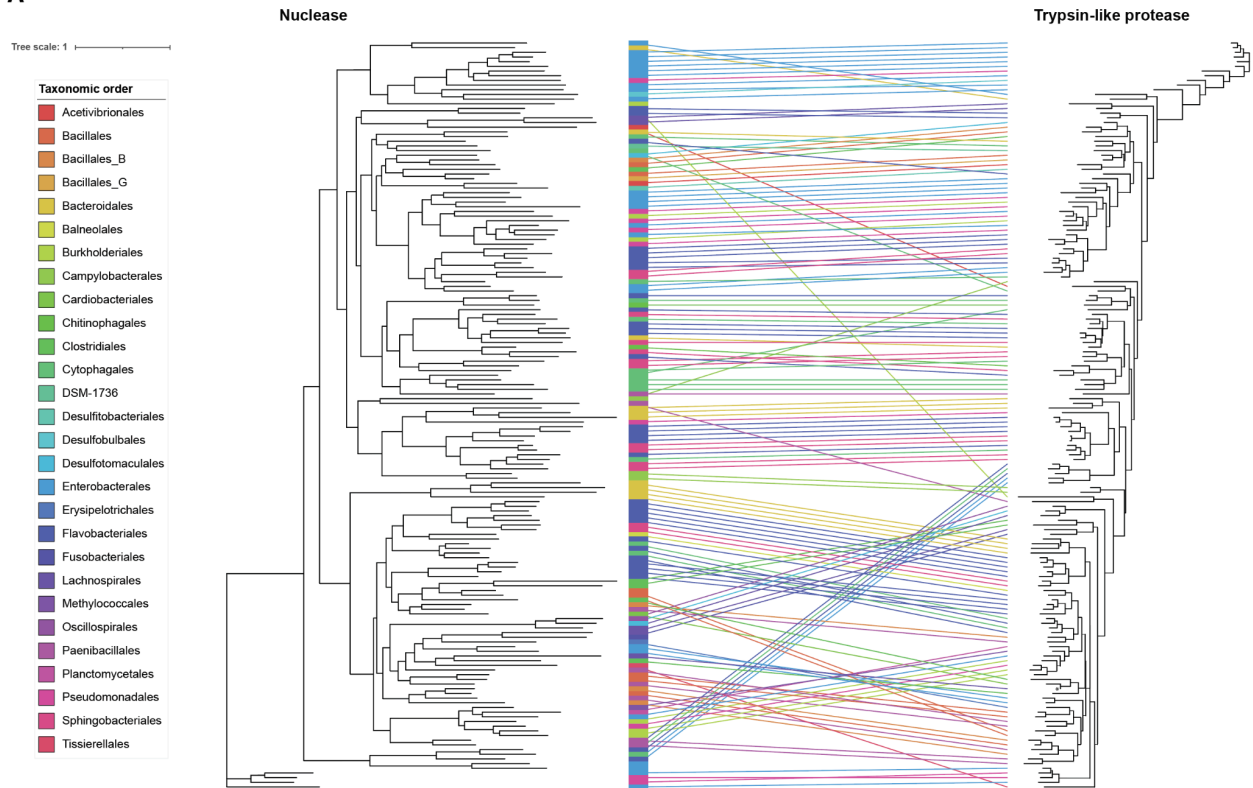
Supplementary Figure 4-2. Modular evolution of defense-associated MBL hydrolase-like nucleases. **A)** Phylogenetic tree of diverse MBL hydrolase-like nuclease sequences in bacteria. The phylogenetic tree is rooted at a large clade containing MBL hydrolase-like nucleases without known protease associations which encode ComE nuclease domains. Tree leaves are colored according to inclusion of the predicted nuclease in known defense systems as determined by PADLOC search of a 20 kbp genomic context. The outer track indicates association with nearby protease families. Three major operonic protease families - subtilase-, caspase-/CHAT- and trypsin-family proteases were detected. Branches with support values < 60 are deleted, resulting in polytomy. Select bootstrap support values are shown. The clade containing trypsin-associated nucleases is shown as a subtree in **Fig. 4-1B**. A subtree of protease-associated nucleases is shown with clades collapsed in **Fig. 4-4A**. **B)** Unrooted form of phylogenetic tree from (A). Colored ranges indicate associated proteases as in (A). **C)** Genomic loci of nucleases from the phylogeny in (A). Select loci from each major protease-associated clade are labelled in A. Core defense genes, proteases, other detected defense systems in the gene neighborhood and genes associated with mobile genetic element mobility are annotated as in (A) and according to the key below.

A

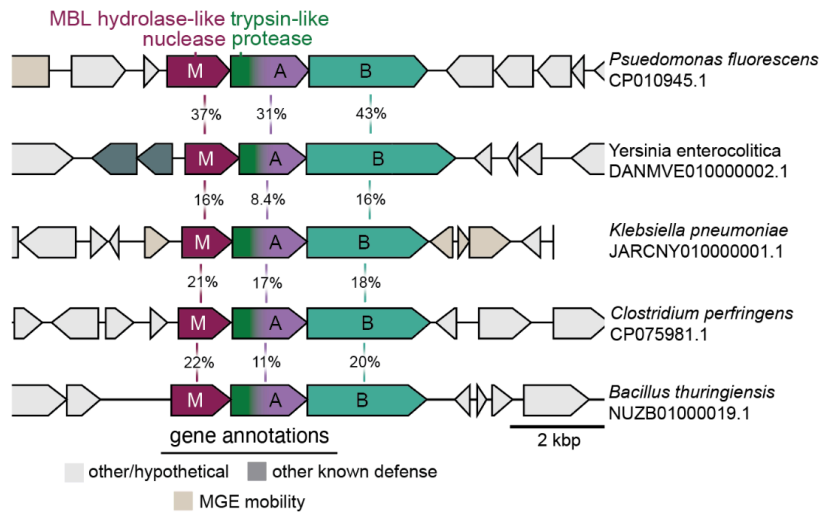


Supplementary Figure 4-3. Comparison between species and nuclease trees and reconciliation analysis. A) Tanglegram comparing a species tree extracted from the Genome Taxonomy Database (GTDB) and select homologs of MBL hydrolase-type nucleases encoded by these species (see Methods). Taxonomic order is indicated by the species tree outer track. Lines connecting leaves show which species encode the corresponding nuclease and serve as a visualization for topological incongruence. Formal reconciliation analyses (GeneRax, UndatedDTL, see Methods) suggest that horizontal transfer and gene loss, but not duplication, primarily account for the evolutionary history of these genes (duplication = $1e-07$, loss = 0.266004, transfer = 0.413985).

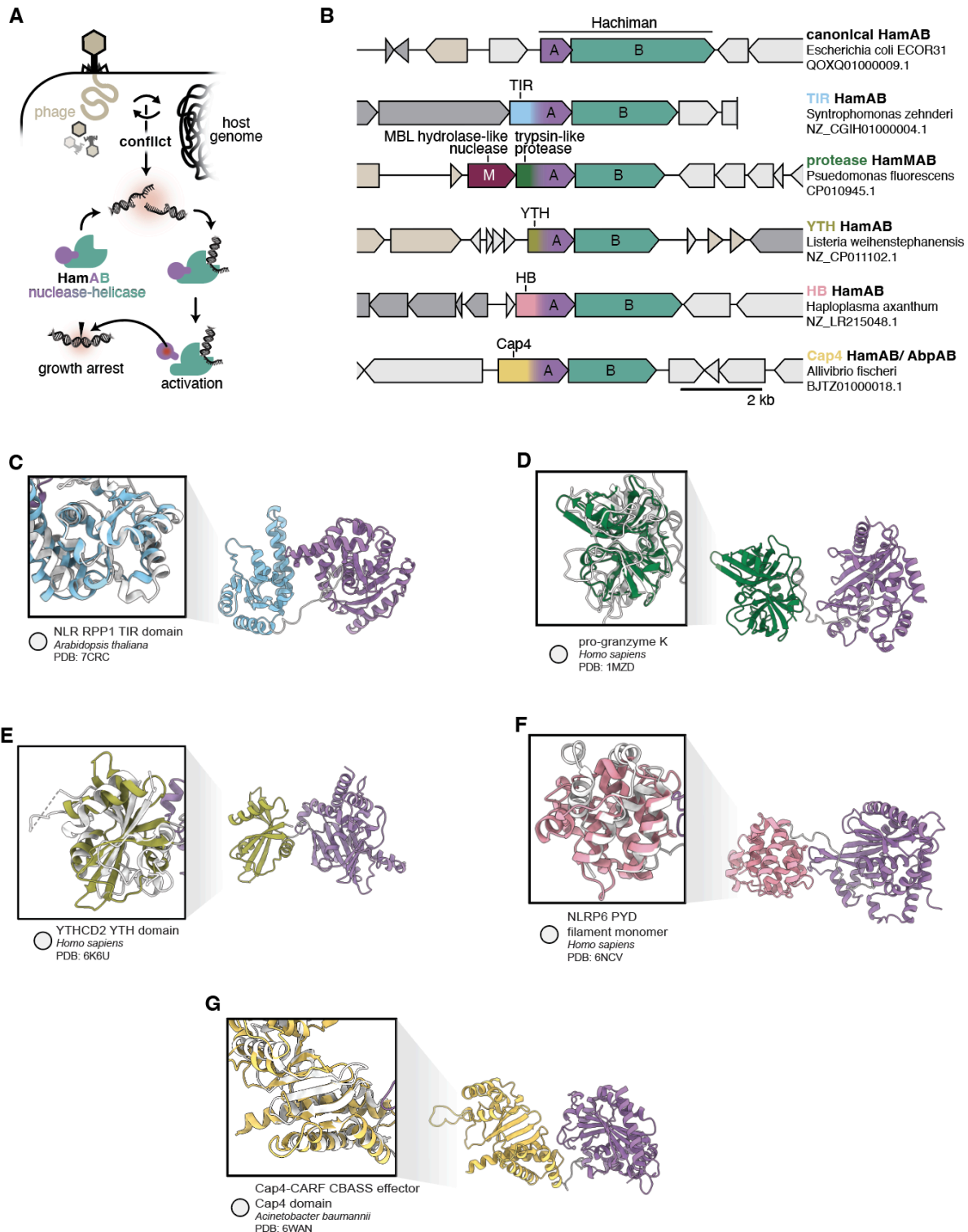
A



Supplementary Figure 4-4. Comparison of nuclease and trypsin-like protease phylogenies. A) Tanglegram comparing phylogeny of select homologs of MBL hydrolase-type nucleases and a phylogeny of the corresponding trypsin-like protease (see Methods). Taxonomic order is indicated by the nuclease tree outer track. Lines are colored according to nuclease taxonomy and connect to co-encoded proteases.

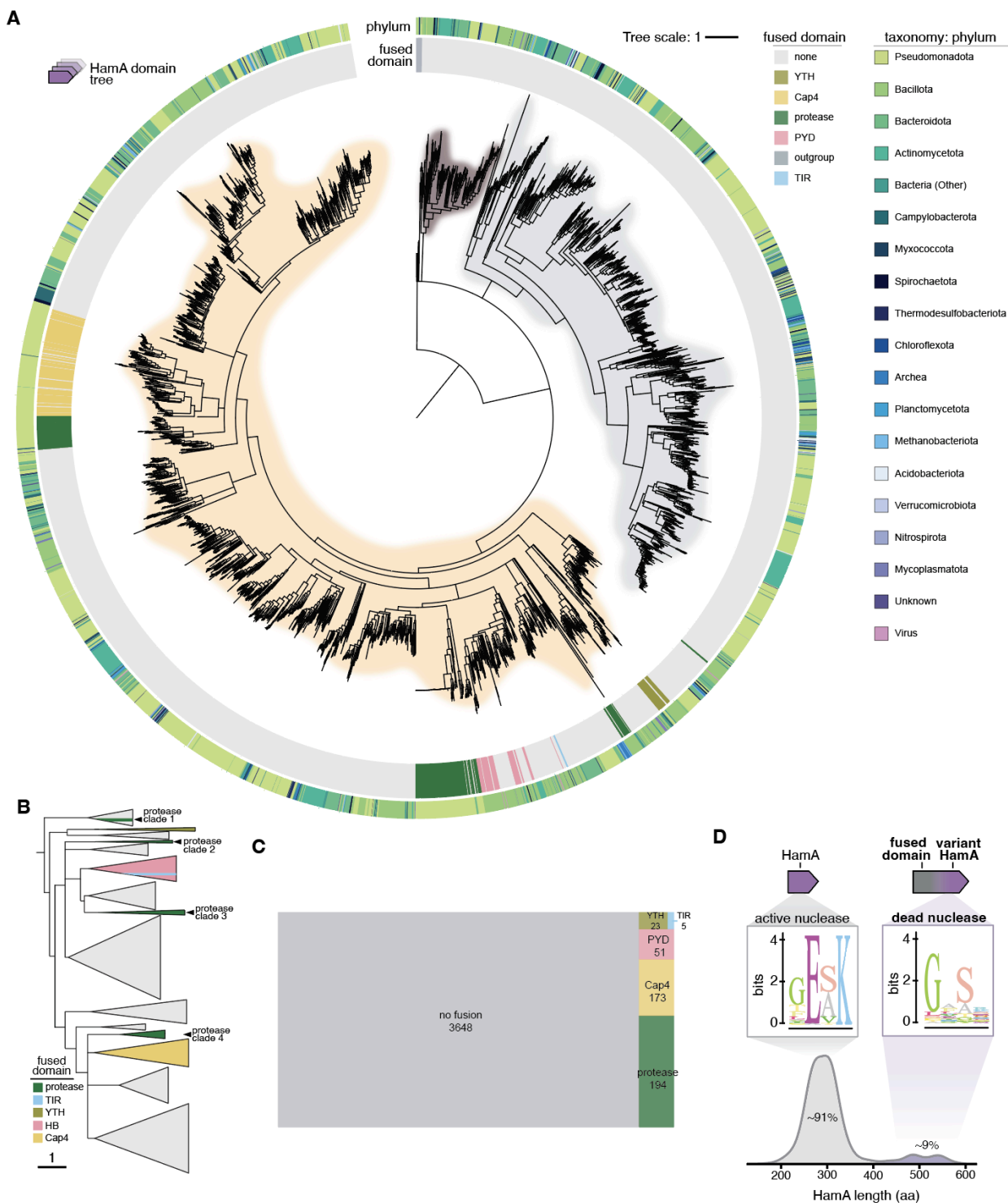


Supplementary Figure 4-5. Nuclease-protease Hachiman systems are highly diverse. Genomic loci encoding nuclease-protease Hachiman systems from different clades of the nuclease phylogeny in **Supplementary Fig. 4-2A**. Percent identity between different genes are shown between loci. *P. fluorescens* and *Y. enterocolitica* sequences originate from the same clade.



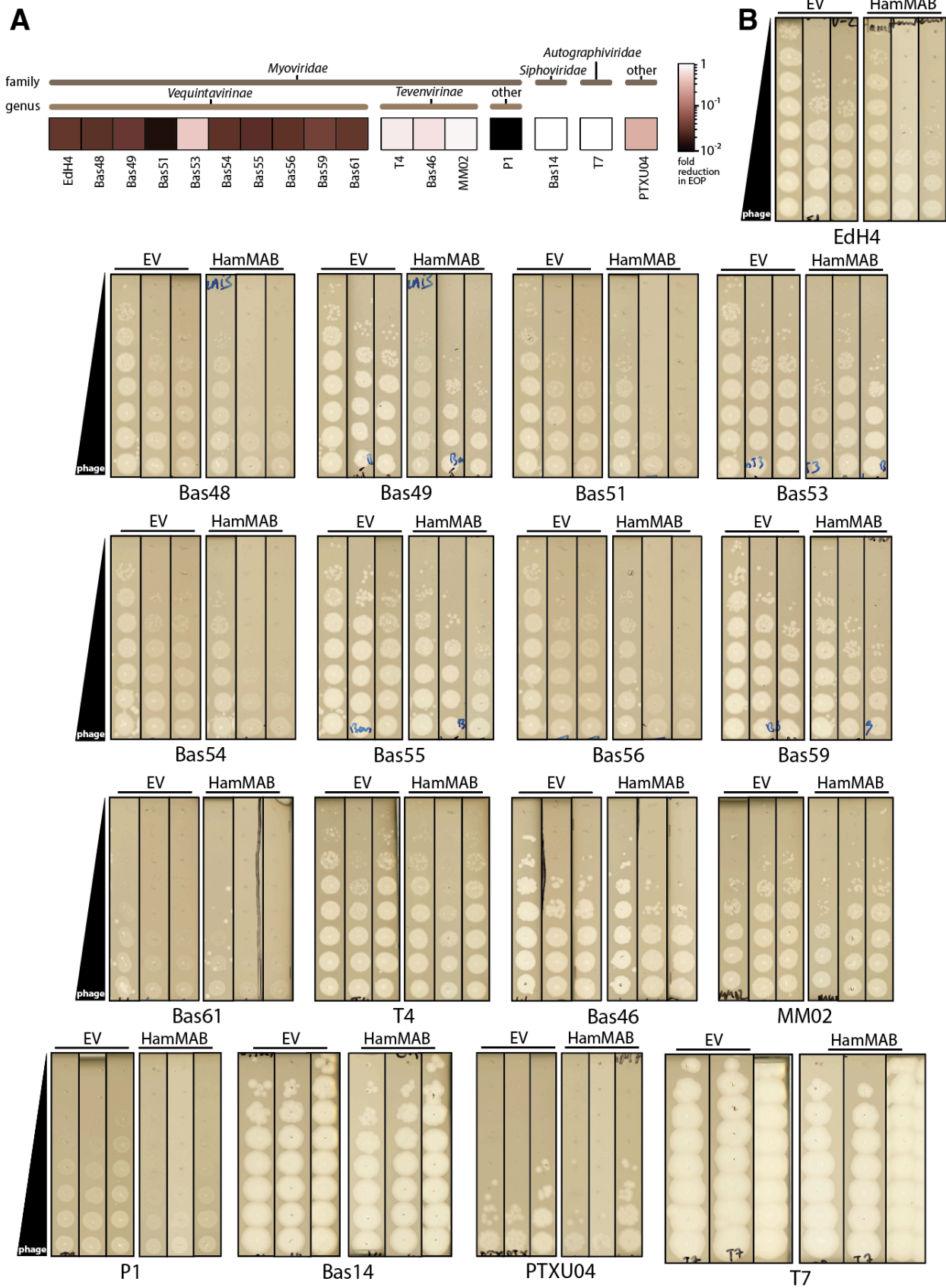
Supplementary Figure 4-6. Hachiman HamA acquired multiple additional domains. A) Model for Hachiman antiphage defense. **B)** Loci illustrating the genetic architecture of canonical and fusion variant Hachiman systems. The five fusion-HamA systems identified are: Toll/interleukin-1 receptor (TIR) domain-HamA, trypsin protease-HamA with upstream metallo- β -lactamase fold hydrolase-like nuclease (HamM), YT521-B homology (YTH) domain-HamA, helical bundle (HB) domain-HamA, and CBASS

activated protein 4 (Cap4) domain-HamA or AbpA. **C-G**) AlphaFold structural predictions of domain-HamA fusions shown in (B). Inset boxes show structural superpositions with DALI-predicted homologous experimental structures, which are shown in white and are indicated below in each case.

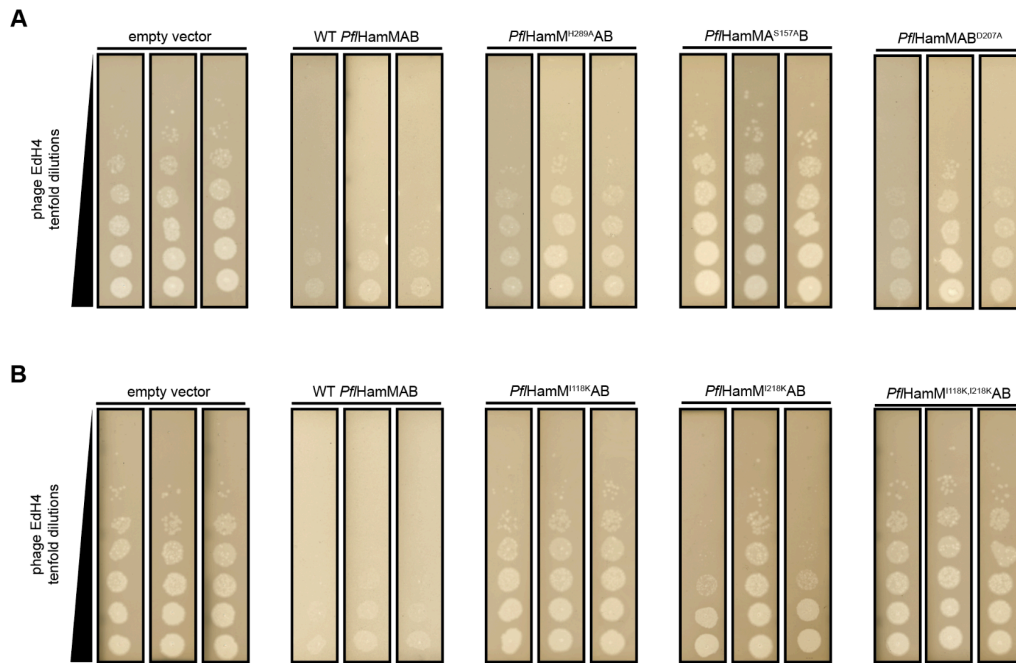


Supplementary Figure 4-7. HamA domain phylogeny and loss of nuclease activity. **A)** Phylogenetic tree of 4,094 HamA domain sequences. The tree is rooted at a clade containing 10 homologs of the *P. aquatilis* type IIS restriction endonuclease domain. The inner track indicates presence of a fusion domain, but these were excluded from phylogenetic analysis to ascertain the phylogeny exclusively of the HamA domain. The outer track indicates associated phyla. Branches with support values < 60 are deleted, resulting in polytomy. **B)** Cartoon of the fusion containing clade in (A), with clades collapsed and resulting in polytomy.

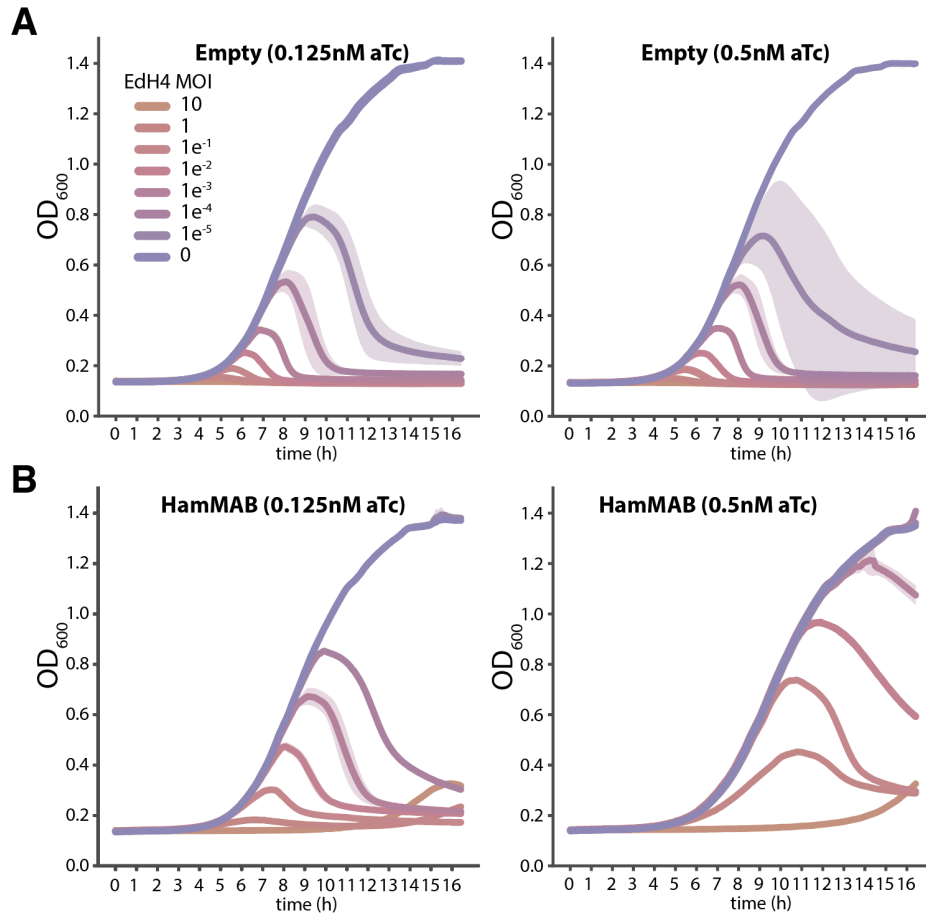
color-coded to indicate fusion content. **C)** Fraction of HamA sequences encoding fusions. **D)** Sequence logos of the nuclease active site for canonical versus domain fusion HamA variants are shown above a density plot showing the distribution of HamA protein lengths across HamA diversity. HamA variants were manually curated from all HamA domain sequences from panel (A) and aligned.



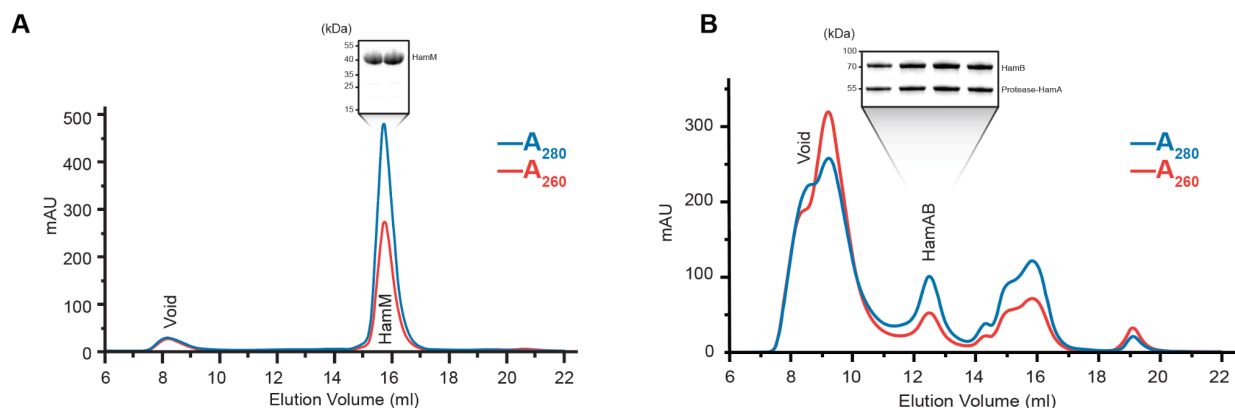
Supplementary Figure 4-8. *Pseudomonas fluorescens* HamMAB is antiphage. A) Heatmap of the efficiency of plating of diverse phages on *E. coli* expressing *P. fluorescens* HamMAB. **B)** Images of triplicate plaque assays where *E. coli* expressing *Pfl*HamMAB was challenged with diverse bacteriophages indicated. Experiments are summarized in (A).



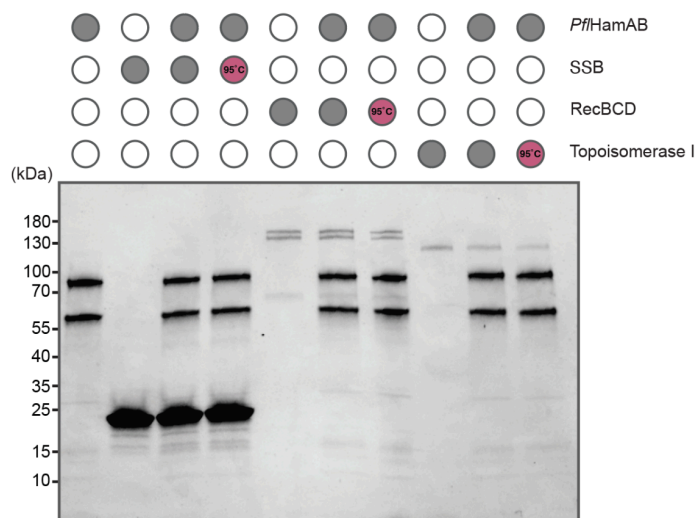
Supplementary Figure 4-9. Plaque assays of *P. fluorescens* HamMAB mutants. A) Images of triplicate plaque assays where *E. coli* expressing *Pfl*HamMAB and catalytic mutants were challenged with tenfold dilutions of phage EdH4. **B)** Images of triplicate plaque assays where *E. coli* expressing *Pfl*HamMAB and cut site mutants of HamM were challenged with tenfold dilutions of phage EdH4.



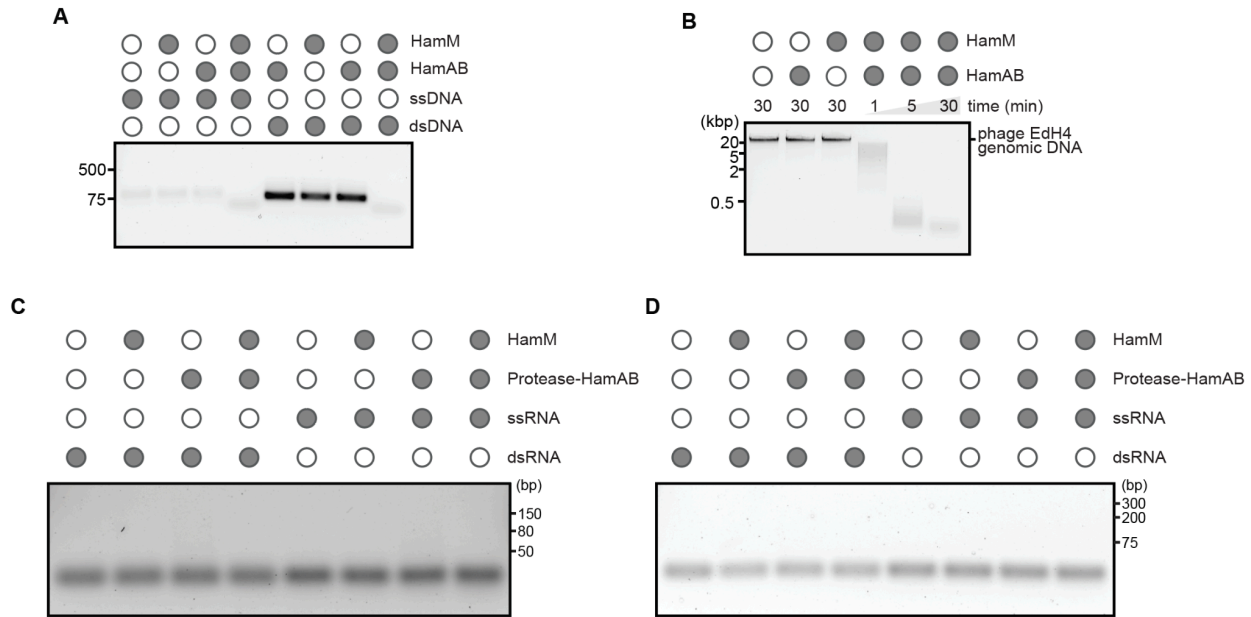
Supplementary Figure 4-10. Complete liquid infection data for PflHamMAB. A) Growth curves of *E. coli* empty vector control during EdH4 infection at the specified multiplicity of infection (MOI) and level of induction. **B)** Growth curves of *E. coli* expressing PflHamMAB during EdH4 infection at the specified multiplicity of infection (MOI) and level of induction. In all cases, data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates.



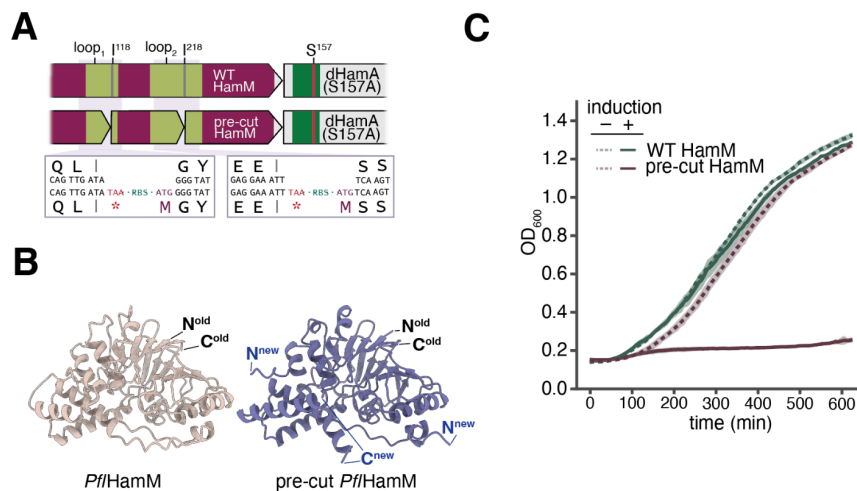
Supplementary Figure 4-11. Purification of HamM and HamAB. **A)** Size exclusion chromatography trace of HamM after TEV protease treatment and removal of MBP by reverse immobilized metal affinity purification (IMAC). Corresponding peaks of elution fractions are run on a Coomassie PAGE gel. **B)** Size exclusion chromatography trace of HamM after TEV protease treatment and removal of MBP by reverse IMAC. Corresponding peaks of elution fractions were run on a Coomassie PAGE gel.



Supplementary Figure 4-12. In vitro incubation of *Pfl*HamAB with Single-Stranded DNA Binding Protein (SSB), RecBCD, and Topoisomerase I. In vitro reactions of HamAB with SSB, RecBCD, and Topoisomerase I visualized on a Coomassie PAGE gel. In lanes 4, 7, and 10 (from left to right), protein substrates were pre-boiled before addition of HamAB.

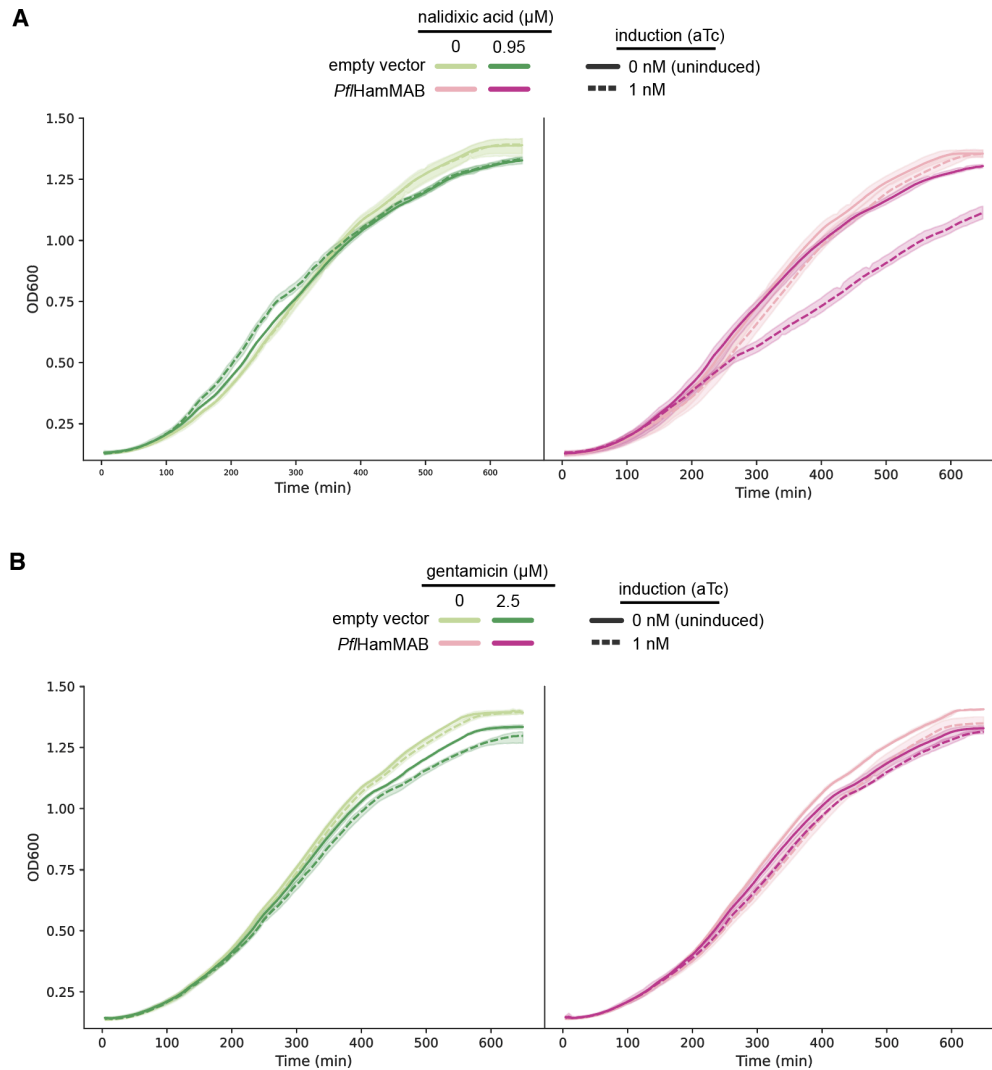


Supplementary Figure 4-13. *In vitro* cleavage of various nucleic acids by *Pfl*HamMAB. **A) *In vitro* reactions of HamM and HamAB against ssDNA and ssDNA, visualized on a 2% agarose gel stained with SYBR Gold. **B)** *In vitro* phage EdH4 gDNA cleavage assay with HamM and HamAB with time course. **C)** *In vitro* reactions of HamM and HamAB against dsRNA and ssRNA, visualized on a 2% agarose gel stained with SYBR Gold. **D)** *In vitro* reactions of HamM and HamAB against dsRNA and ssRNA in buffer supplemented with 1 mM ZnCl₂, visualized on a 2% agarose gel stained with SYBR Gold.**

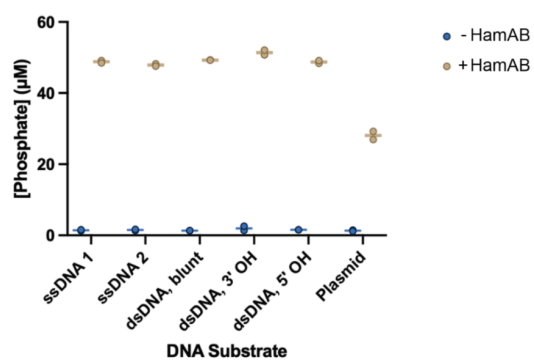


Supplementary Figure 4-14. Pre-cut HamM is constitutively toxic to host cells. **A) Diagram of comparing portions of the wildtype *Pfl*HamMAB locus with the 'pre-cut' version of the HamM in HamaA protease-dead backgrounds. DNA and resulting protein sequences are shown as insets below loop 1 and loop 2 proteolysis insertion regions to illustrate how breaks in the peptide backbone were programmed. **B)** AlphaFold3 predicted structures of wildtype *Pfl*HamMAB and 'pre-cut' tripeptide. The locations of the new N- and C-termini are shown. **C)** Growth curves of *E. coli* expressing either native**

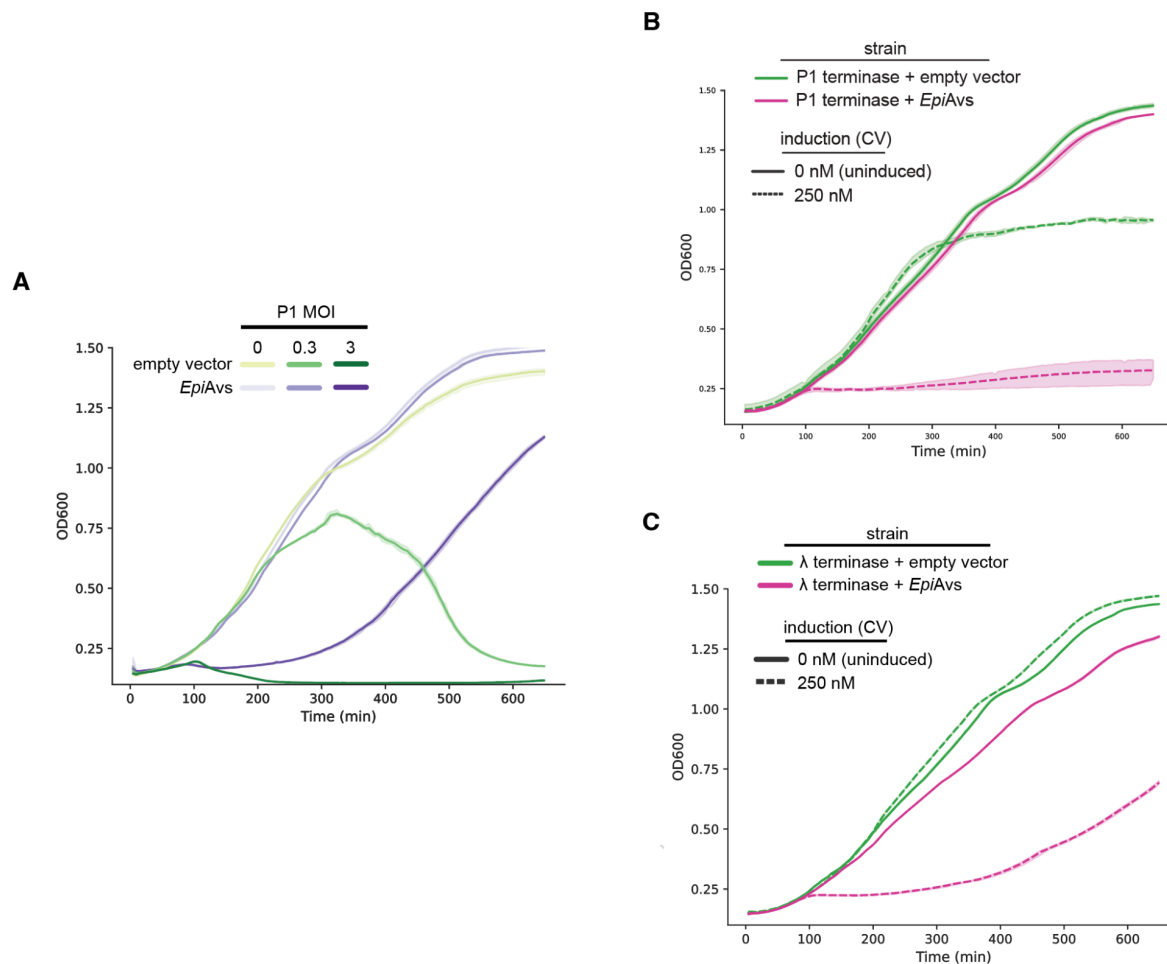
*Pfl*HamMAB or 'pre-cut' *Pfl*HamMAB, each with a deactivated protease domain (S157A), with or without induction. Data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates.



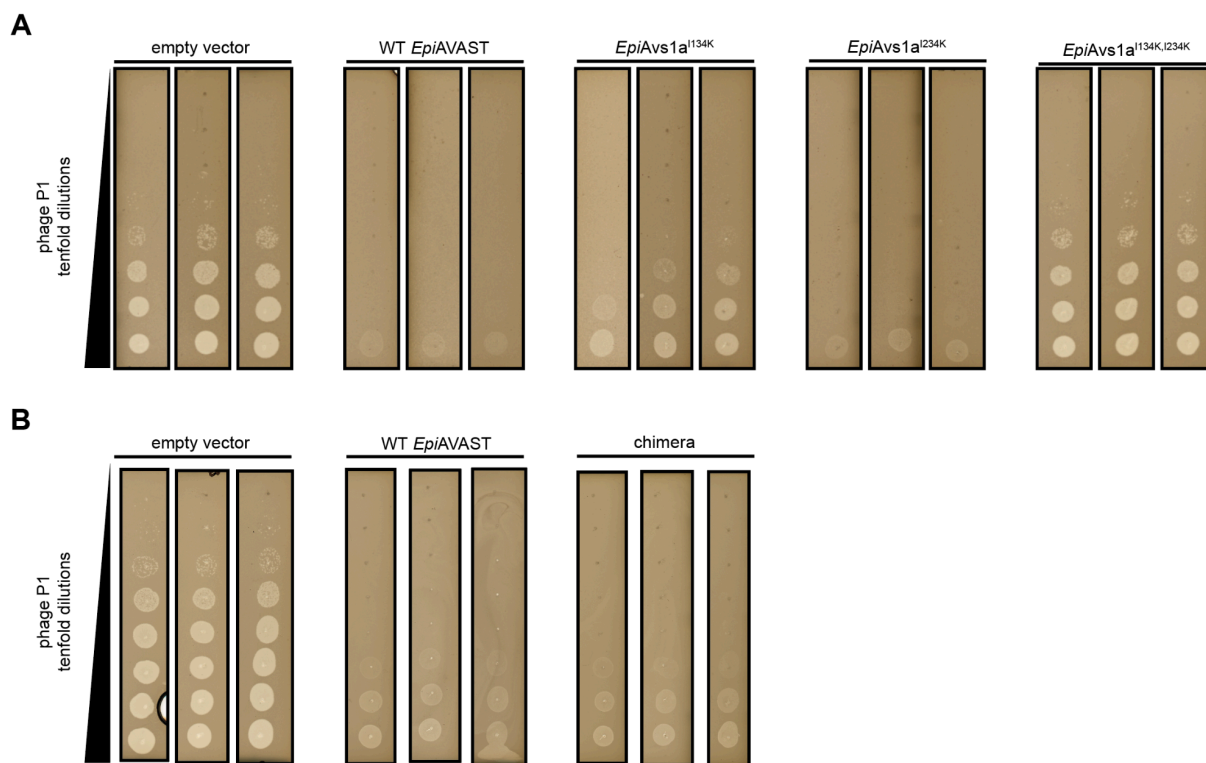
Supplementary Figure 4-15. Drug-induced *Pfl*HamMAB toxicity. **A)** Growth curves of *E. coli* expressing either native *Pfl*HamMAB or an empty vector control, with or without the indicated concentrations of nalidixic acid and inducer (anhydrotetracycline). **B)** Growth curves of *E. coli* expressing either native *Pfl*HamMAB or an empty vector control, With or without the indicated concentrations of gentamicin and inducer (anhydrotetracycline). In all cases, data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates.



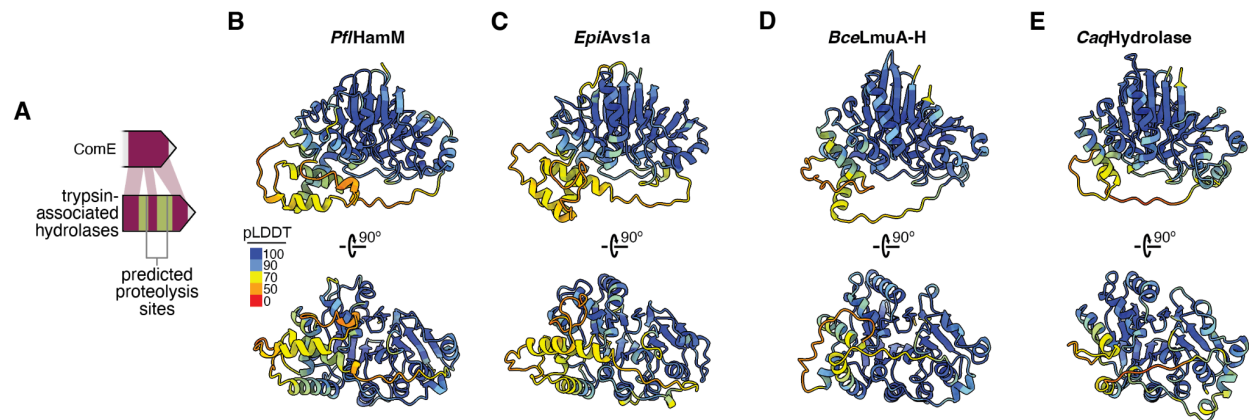
Supplementary Figure 4-16. *In vitro* ATPase assay of *Pfl*HamAB. Malachite green ATPase assays of HamAB with ATP against a panel of DNA substrates. Individual data points of two independent biological replicates are shown.



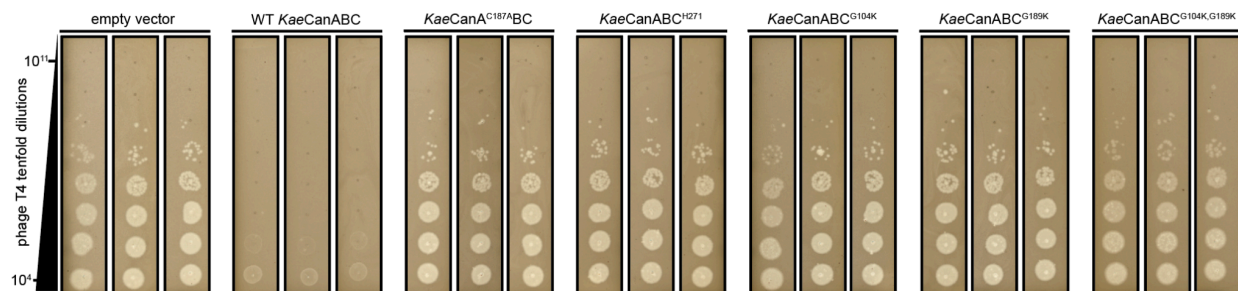
Supplementary Figure 4-17. *EpiAvs* liquid infection and terminase-mediated activation. A) Growth curves of *E. coli* empty vector control or expressing *EpiAvs* during phage P1 infection at the specified multiplicity of infection (MOI). **B)** Growth curves of *E. coli* coexpressing phage P1 terminase gene under control of the pJEX system with either *EpiAvs* or an empty vector control. Induction of the terminase gene with crystal violet (CV) is indicated with a dashed line. **C)** Growth curves of *E. coli* coexpressing phage lambda terminase gene under control of the pJEX system with either *EpiAvs* or an empty vector control. Induction of the terminase gene with crystal violet (CV) is indicated with a dashed line. In every case, data are shown as mean $\pm$ standard deviation (shaded area) across three independent biological replicates.



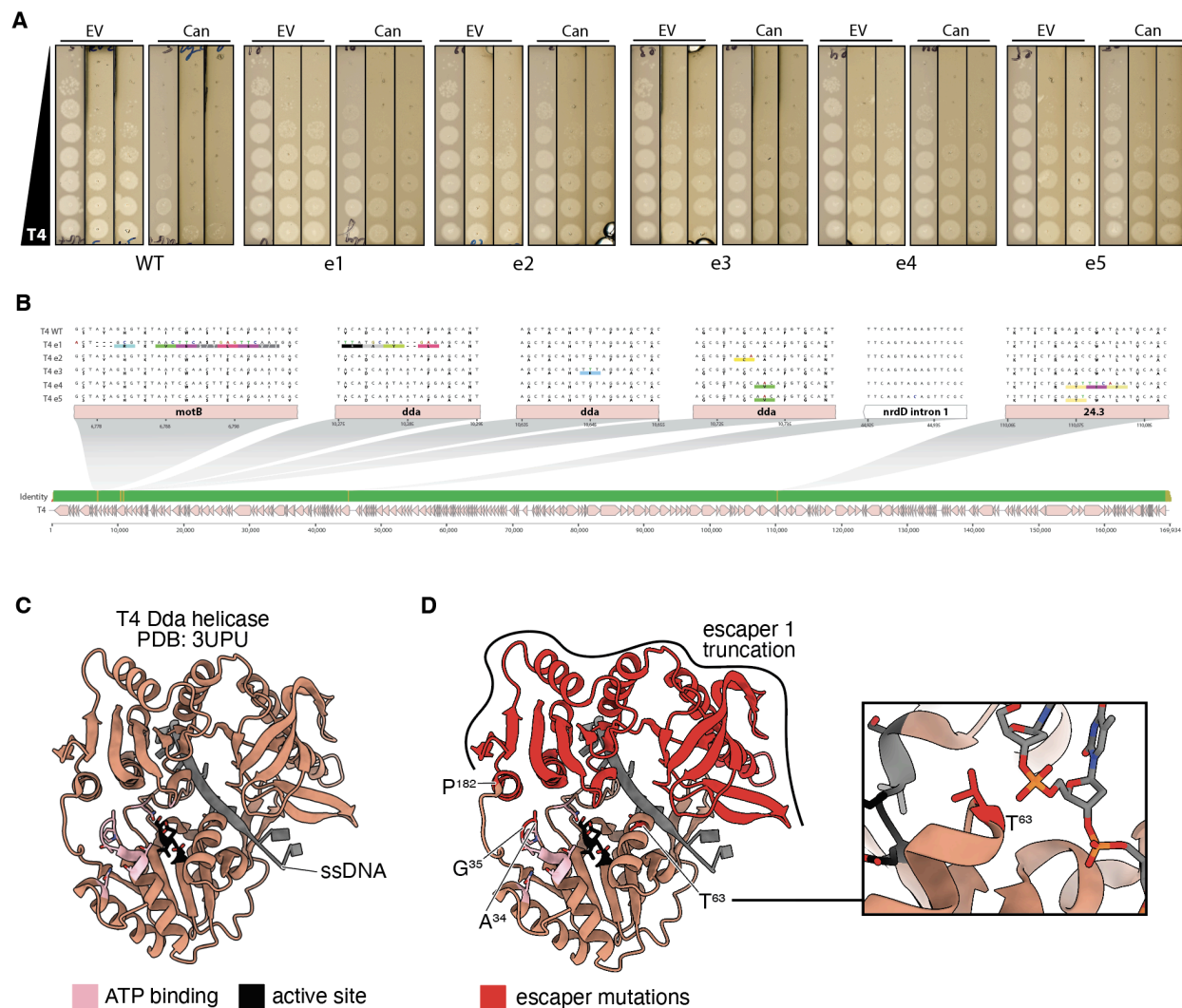
Supplementary Figure 4-18. Plaque assays of *E. piriflorinigrans* AVAST and Avs1a mutants. A) Images of triplicate plaque assays where *E. coli* expressing *EpiAVAST* and catalytic mutants of Avs1a were challenged with tenfold dilutions of phage P1. **B)** Images of triplicate plaque assays where *E. coli* expressing *EpiAVAST* and a chimeric *EpiAvs* system with an Avs1a nuclease from *Citrobacter* sp. ESBL3 (NZ_JAZEVS010000002.1).



Supplementary Figure 4-19. Structural comparisons of trypsin-associated nucleases. **A)** Model for the emergence of autoinhibitory ‘gating’ insertions (green) in trypsin-associated nucleases in antiphage systems. **B-E)** AlphaFold3 predicted protein structures for nuclease genes in Hachiman, AVAST, Lamassu and DRT (UG9) systems, as in **Supplementary Fig. 4-1** but with an additional view shown and colored by pLDDT. In each case, protein structures were predicted using loci from **Fig. 4-1A**. Predicted structural snapshots are aligned to show homology.



Supplementary Figure 4-20. Plaque assays of *K. aerogenes* CanABC and mutants. **A)** Images of triplicate plaque assays where *E. coli* expressing KaeCanABC and various catalytic and cut site mutants were challenged with tenfold dilutions of phage T4.



Supplementary Figure 4-21. Phage escape of *KaeCanABC* involves mutation of the *dda* helicase.
A) Images of triplicate plaque assays where *E. coli* expressing *K. aerogenes* CanABC was challenged with 10-fold dilutions of WT and isolated escaper phage T4. **B)** Overview of the whole-genome alignments of WT and Canu-escaping T4 phages. **C)** Overview of the T4 Dda helicase structure (PDB:3UPU) with ssDNA, with ATP binding and active site residues labeled. **D)** Location of phage T4 escape mutants on the structure from (C). Locations of point mutations and truncations are shown in red, as in a close-up of residue T63 to demonstrate proximity to DNA.

References

1. Metchnikoff, E. (1887). Sur la lutte des cellules de l'organisme contre l'invasion des microbes. *Ann. Inst. Pasteur*.
2. Janeway, C.A., Jr (1992). The immune system evolved to discriminate infectious nonself from noninfectious self. *Immunol. Today* 13, 11–16.
3. Janeway, C.A., Jr, and Medzhitov, R. (2002). Innate immune recognition. *Annu. Rev. Immunol.* 20, 197–216.
4. Fitzgerald, K.A., and Kagan, J.C. (2020). Toll-like receptors and the control of immunity. *Cell* 180, 1044–1066.
5. Chow, K.T., Gale, M., Jr, and Loo, Y.-M. (2018). RIG-I and Other RNA Sensors in Antiviral Immunity. *Annu. Rev. Immunol.* 36, 667–694.
6. Yu, P., Zhang, X., Liu, N., Tang, L., Peng, C., and Chen, X. (2021). Pyroptosis: mechanisms and diseases. *Signal Transduct. Target. Ther.* 6, 128.
7. Lopatina, A., Tal, N., and Sorek, R. (2020). Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annu Rev Virol* 7, 371–384.
8. Unterholzner, L., Sumner, R.P., Baran, M., Ren, H., Mansur, D.S., Bourke, N.M., Randow, F., Smith, G.L., and Bowie, A.G. (2011). Vaccinia virus protein C6 is a virulence factor that binds TBK-1 adaptor proteins and inhibits activation of IRF3 and IRF7. *PLoS Pathog.* 7, e1002247.
9. Kärre, K., Ljunggren, H.G., Piontek, G., and Kiessling, R. (1986). Selective rejection of H-2-deficient lymphoma variants suggests alternative immune defence strategy. *Nature* 319, 675–678.
10. Matzinger, P. (1994). Tolerance, danger and the extended family. *Annu. Rev. Immunol.* 12, 991–1045.
11. Tang, D., Kang, R., Coyne, C.B., Zeh, H.J., and Lotze, M.T. (2012). PAMPs and DAMPs: signal 0s that spur autophagy and immunity. *Immunol. Rev.* 249, 158–175.
12. Dinarello, C.A. (2018). Overview of the IL-1 family in innate inflammation and acquired immunity. *Immunol. Rev.* 281, 8–27.
13. Van, E.W., Spiering, R., Broere, F., and R., van der Z. (2012). A case of mistaken identity: HSPs are no DAMPs but DAMPERs. *Cell Stress and Chaperones* 17, 281–292.
14. Rosenblum, M.D., Remedios, K.A., and Abbas, A.K. (2015). Mechanisms of human autoimmunity. *J. Clin. Invest.* 125, 2228–2233.

15. Deighton, C. (1993). Beta haemolytic streptococci and reactive arthritis in adults. *Ann. Rheum. Dis.* 52, 475–482.
16. Gonzalez, S., González-Rodríguez, A.P., Suárez-Álvarez, B., López-Soto, A., Huergo-Zapico, L., and Lopez-Larrea, C. (2011). Conceptual aspects of self and nonself discrimination. *Self Nonself* 2, 19–25.
17. Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214.
18. Kau, A.L., Ahern, P.P., Griffin, N.W., Goodman, A.L., and Gordon, J.I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336.
19. Seekatz, A.M., Safdar, N., and Khanna, S. (2022). The role of the gut microbiome in colonization resistance and recurrent *Clostridioides difficile* infection. *Therap. Adv. Gastroenterol.* 15, 17562848221134396.
20. Prak, E.T., and Kazazian, H.H., Jr (2000). Mobile elements and the human genome. *Nat. Rev. Genet.* 1, 134–144.
21. Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403, 785–789.
22. Gerdes, P., Karlsson, O., Garza, R., Atacho, D.A.M., Hsieh, P., Prajapati, B., Muralidharan, C., Adami, A., Davis-Hansson, C., Johansson, E.M., et al. (2025). Retroviral insertions contributed to the divergence of human and chimpanzee brains. *bioRxiv*org. <https://doi.org/10.64898/2025.12.12.693858>.
23. Pancer, Z., and Cooper, M.D. (2006). The evolution of adaptive immunity. *Annu. Rev. Immunol.* 24, 497–518.
24. Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature* 381, 751–758.
25. Sakaguchi, S., Yamaguchi, T., Nomura, T., and Ono, M. (2008). Regulatory T cells and immune tolerance. *Cell* 133, 775–787.
26. Kelley, J., de Bono, B., and Trowsdale, J. (2005). IRIS: a database surveying known human immune system genes. *Genomics* 85, 503–511.
27. Mordret, E., Hervé, A., Tesson, F., Vaysset, H., Clabby, T., Loubat, A., Shomar, H., Planel, R., Lavenir, R., Cury, J., et al. (2026). Protein and genomic language models uncover the unexplored diversity of bacterial immunity. *Science* 392, eadv8275.
28. Bernheim, A., Cury, J., and Poirier, E.Z. (2024). The immune modules conserved across the tree of life: Towards a definition of ancestral immunity. *PLoS Biol.* 22, e3002717.
29. Tal, N., and Sorek, R. (2022). SnapShot: Bacterial immunity. *Cell* 185, 578–578.e1.

30. Ledvina, H.E., and Whiteley, A.T. (2024). Conservation and similarity of bacterial and eukaryotic innate immunity. *Nat. Rev. Microbiol.*, 1–15.
31. Monod, J. (1971). Chance and necessity.
32. Luria, S.E., and Human, M.L. (1952). A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.* 64, 557–569.
33. Nathans, D., and Smith, H.O. (1975). Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu. Rev. Biochem.* 44, 273–293.
34. Tyson, G.W., and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* 10, 200–207.
35. Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015). Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* 527, 535–538.
36. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712.
37. Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960–964.
38. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821.
39. Unterholzner, S.J., Poppenberger, B., and Rozhon, W. (2013). Toxin-antitoxin systems: Biology, identification, and application: Biology, identification, and application. *Mob. Genet. Elements* 3, e26219.
40. Aravind, L., Dixit, V.M., and Koonin, E.V. (1999). The domains of death: evolution of the apoptosis machinery. *Trends Biochem. Sci.* 24, 47–53.
41. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359. <https://doi.org/10.1126/science.aar4120>.
42. Georjon, H., and Bernheim, A. (2023). The highly diverse antiphage defence systems of bacteria. *Nat. Rev. Microbiol.* <https://doi.org/10.1038/s41579-023-00934-x>.
43. Gao, L., Altae-Tran, H., Böhning, F., Makarova, K.S., Segel, M., Schmid-Burgk, J.L., Koob, J., Wolf, Y.I., Koonin, E.V., and Zhang, F. (2020). Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* 369, 1077–1084.
44. Gao, L.A., Wilkinson, M.E., Strecker, J., Makarova, K.S., Macrae, R.K., Koonin, E.V., and Zhang, F. (2022). Prokaryotic innate immunity through pattern recognition of conserved

viral proteins. *Science* 377, eabm4096.

45. Kibby, E.M., Conte, A.N., Burroughs, A.M., Nagy, T.A., Vargas, J.A., Whalen, L.A., Aravind, L., and Whiteley, A.T. (2023). Bacterial NLR-related proteins protect against phage. *Cell* 186, 2410–2424.e18.
46. Ofir, G., Herbst, E., Baroz, M., Cohen, D., Millman, A., Doron, S., Tal, N., Malheiro, D.B.A., Malitsky, S., Amitai, G., et al. (2021). Antiviral activity of bacterial TIR domains via immune signalling molecules. *Nature* 600, 116–120.
47. Whiteley, A.T., Eaglesham, J.B., de Oliveira Mann, C.C., Morehouse, B.R., Lowey, B., Nieminen, E.A., Danilchanka, O., King, D.S., Lee, A.S.Y., Mekalanos, J.J., et al. (2019). Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. *Nature* 567, 194–199.
48. Morehouse, B.R., Govande, A.A., Millman, A., Keszei, A.F.A., Lowey, B., Ofir, G., Shao, S., Sorek, R., and Kranzusch, P.J. (2020). STING cyclic dinucleotide sensing originated in bacteria. *Nature* 586, 429–433.
49. Moody, E.R.R., Álvarez-Carretero, S., Mahendrarajah, T.A., Clark, J.W., Betts, H.C., Dombrowski, N., Szánthó, L.L., Boyle, R.A., Daines, S., Chen, X., et al. (2024). The nature of the last universal common ancestor and its impact on the early Earth system. *Nat. Ecol. Evol.*, 1–13.
50. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.
51. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42, 243–246.
52. Mushegian, A.R. (2020). Are there 10³¹ virus particles on Earth, or more, or fewer? *J. Bacteriol.* 202, e00052–20.
53. Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812.
54. Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. *Bioessays* 33, 43–51.
55. Lopez, J.A., McKeithen-Mead, S., Shi, H., Nguyen, T.H., Huang, K.C., and Good, B.H. (2025). Abundance measurements reveal the balance between lysis and lysogeny in the human gut microbiome. *Current Biology* 35, 2282–2294.e11.
56. Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far?: Prophage genomics. *Mol. Microbiol.* 49, 277–300.
57. Beavogui, A., Lacroix, A., Wiart, N., Poulain, J., Delmont, T.O., Paoli, L., Wincker, P., and Oliveira, P.H. (2024). The defensome of complex bacterial communities. *Nat. Commun.* 15,

58. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* *13*, 2561.
59. Mestre, M.R., González-Delgado, A., Gutiérrez-Rus, L.I., Martínez-Abarca, F., and Toro, N. (2020). Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res.* *48*, 12632–12647.
60. Millman, A., Melamed, S., Amitai, G., and Sorek, R. (2020). Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat. Microbiol.* *5*, 1608–1615.
61. Rousset, F., and Sorek, R. (2023). The evolutionary success of regulated cell death in bacterial immunity. *Curr. Opin. Microbiol.* *74*, 102312.
62. Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* *26*, 335–340.
63. Aframian, N., Omer Bendori, S., Hen, T., Guler, P., and Eldar, A. (2025). Expression level of anti-phage defence systems controls a trade-off between protection range and autoimmunity. *Nat. Microbiol.* *10*, 1954–1962.
64. Koonin, E.V., Makarova, K.S., and Wolf, Y.I. (2017). Evolutionary genomics of defense systems in Archaea and bacteria. *Annu. Rev. Microbiol.* *71*, 233–261.
65. Hussain, F.A., Dubert, J., Elsherbini, J., Murphy, M., VanInsberghe, D., Arevalo, P., Kauffman, K., Rodino-Janeiro, B.K., Gavin, H., Gomez, A., et al. (2021). Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* *374*, 488–492.
66. Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* *193*, 6039–6056.
67. Bernheim, A., and Sorek, R. (2020). The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* *18*, 113–119.
68. Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* *40*, 5569–5576.
69. Nuñez, J.K., Lee, A.S.Y., Engelman, A., and Doudna, J.A. (2015). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* *519*, 193–198.
70. Marraffini, L.A., and Sontheimer, E.J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* *463*, 568–571.
71. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system.

Microbiology 155, 733–740.

72. Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S., and Koonin, E.V. (2017). Phylogenomics of Cas4 family nucleases. *BMC Evol. Biol.* 17, 232.
73. Hu, C., Almendros, C., Nam, K.H., Costa, A.R., Vink, J.N.A., Haagsma, A.C., Bagde, S.R., Brouns, S.J.J., and Ke, A. (2021). Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* 598, 515–520.
74. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* 18, 67–83.
75. Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of pre-spacers for precise CRISPR adaptation. *Nature* 579, 141–145.
76. Ramachandran, A., Summerville, L., Learn, B.A., DeBell, L., and Bailey, S. (2020). Processing and integration of functionally oriented pre-spacers in the *Escherichia coli* CRISPR system depends on bacterial host exonucleases. *J. Biol. Chem.* 295, 3403–3414.
77. Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* 8, 317–327.
78. Sternberg, S.H., Richter, H., Charpentier, E., and Qimron, U. (2016). Adaptation in CRISPR–Cas systems. *Molecular Cell* 61, 797–808.
79. McGinn, J., and Marraffini, L.A. (2019). Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat. Rev. Microbiol.* 17, 7–12.
80. Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). 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.
81. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution* 60, 174–182.
82. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151, 2551–2561.
83. Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468, 67–71.
84. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* 139,

945–956.

85. Wright, A.V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. *Nat. Struct. Mol. Biol.* 23, 876–883.
86. Xiao, Y., Ng, S., Nam, K.H., and Ke, A. (2017). How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* 550, 137–141.
87. Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR–Cas systems. *Cell* 163, 840–853.
88. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390–1400.
89. Gleditsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019). PAM identification by CRISPR–Cas effector complexes: diversified mechanisms and structures. *RNA Biol.* 16, 504–517.
90. Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell Rep.* 22, 3377–3384.
91. Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays. *Mol. Cell* 70, 48–59.e5.
92. Drabavicius, G., Sinkunas, T., Silanskas, A., Gasiunas, G., Venclovas, Č., and Siksnys, V. (2018). DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR–Cas system. *EMBO Rep.* 19. <https://doi.org/10.15252/embr.201745543>.
93. Lee, H., Dhingra, Y., and Sashital, D.G. (2019). The Cas4–Cas1–Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* 8. <https://doi.org/10.7554/eLife.44248>.
94. Dhingra, Y., Suresh, S.K., Juneja, P., and Sashital, D.G. (2022). PAM binding ensures orientational integration during Cas4–Cas1–Cas2-mediated CRISPR adaptation. *Mol. Cell* 82, 4353–4367.e6.
95. Shiriaeva, A.A., Kuznedelov, K., Fedorov, I., Musharova, O., Khvostikov, T., Tsoy, Y., Kurilovich, E., Smith, G.R., Semenova, E., and Severinov, K. (2022). Host nucleases generate prespacers for primed adaptation in the *E. coli* type I-E CRISPR–Cas system. *Sci Adv* 8, eabn8650.
96. Makarova, K.S., and Koonin, E.V. (2015). Annotation and classification of CRISPR–Cas systems. *Methods in Molecular Biology*, 47–75.
97. Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas

- adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534.
98. Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118.
 99. Sasnauskas, G., and Siksnys, V. (2020). CRISPR adaptation from a structural perspective. *Curr. Opin. Struct. Biol.* **65**, 17–25.
 100. Punjani, A., and Fleet, D.J. (2021). 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702.
 101. Jakhanwal, S., Cress, B.F., Maguin, P., Lobba, M.J., Marraffini, L.A., and Doudna, J.A. (2021). A CRISPR-Cas9–integrase complex generates precise DNA fragments for genome integration. *Nucleic Acids Res.* **49**, 3546–3556.
 102. Wright, A.V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and applications of CRISPR systems: harnessing nature’s toolbox for genome engineering. *Cell* **164**, 29–44.
 103. Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR immunological memory requires a host factor for specificity. *Mol. Cell* **62**, 824–833.
 104. Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510.
 105. Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642–671, Table of Contents.
 106. Dillingham, M.S., Spies, M., and Kowalczykowski, S.C. (2003). RecBCD enzyme is a bipolar DNA helicase. *Nature* **423**, 893–897.
 107. Budhathoki, J.B., Xiao, Y., Schuler, G., Hu, C., Cheng, A., Ding, F., and Ke, A. (2020). Real-time observation of CRISPR spacer acquisition by Cas1-Cas2 integrase. *Nat. Struct. Mol. Biol.* **27**, 489–499.
 108. Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175.
 109. Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2017). CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349.
 110. Bhattarai-Kline, S., Lear, S.K., Fishman, C.B., Lopez, S.C., Lockshin, E.R., Schubert, M.G., Nivala, J., Church, G.M., and Shipman, S.L. (2022). Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* **608**, 217–225.
 111. Musharova, O., Klimuk, E., Datsenko, K.A., Metlitskaya, A., Logacheva, M., Semenova, E., Severinov, K., and Savitskaya, E. (2017). Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. *Nucleic Acids Res.* **45**, 3297–3307.

112. Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., et al. (2019). Detection of spacer precursors formed in vivo during primed CRISPR adaptation. *Nat. Commun.* 10, 4603.
113. Zabradý, K., Zabradý, M., Kolesar, P., Li, A.W.H., and Doherty, A.J. (2021). CRISPR-Associated Primase-Polymerases are implicated in prokaryotic CRISPR-Cas adaptation. *Nat. Commun.* 12, 3690.
114. Roman, L.J., and Kowalczykowski, S.C. (1989). Characterization of the helicase activity of the *Escherichia coli* RecBCD enzyme using a novel helicase assay. *Biochemistry* 28, 2863–2873.
115. Roman, L.J., and Kowalczykowski, S.C. (1989). Formation of heteroduplex DNA promoted by the combined activities of *Escherichia coli* recA and recBCD proteins. *J. Biol. Chem.* 264, 18340–18348.
116. Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14, 290–296.
117. Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A.J., and Berger, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* 16, 1153–1160.
118. Punjani, A., Zhang, H., and Fleet, D.J. (2020). Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* 17, 1214–1221.
119. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 30, 70–82.
120. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* 66, 486–501.
121. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* 75, 861–877.
122. Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., and Brown, C.M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* 17, 356.
123. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 473.
124. Tuck, O.T., Adler, B.A., Armbruster, E.G., Lahiri, A., Hu, J.J., Zhou, J., Pogliano, J., and Doudna, J.A. (2024). Genome integrity sensing by the broad-spectrum Hachiman

- antiphage defense complex. *Cell* 187, 6914–6928.e20.
- 125.Singleton, M.R., Dillingham, M.S., and Wigley, D.B. (2007). Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* 76, 23–50.
 - 126.Pyle, A.M. (2008). Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.* 37, 317–336.
 - 127.Bleichert, F., Botchan, M.R., and Berger, J.M. (2017). Mechanisms for initiating cellular DNA replication. *Science* 355. <https://doi.org/10.1126/science.aah6317>.
 - 128.Brosh, R.M., Jr (2013). DNA helicases involved in DNA repair and their roles in cancer. *Nat. Rev. Cancer* 13, 542–558.
 - 129.Fuller-Pace, F.V. (2006). DExD/H box RNA helicases: multifunctional proteins with important roles in transcriptional regulation. *Nucleic Acids Res.* 34, 4206–4215.
 - 130.Jarmoskaite, I., and Russell, R. (2014). RNA helicase proteins as chaperones and remodelers. *Annu. Rev. Biochem.* 83, 697–725.
 - 131.Morton, D.J., Kuiper, E.G., Jones, S.K., Leung, S.W., Corbett, A.H., and Fasken, M.B. (2018). The RNA exosome and RNA exosome-linked disease. *RNA* 24, 127–142.
 - 132.Martin, R., Straub, A.U., Doebele, C., and Bohnsack, M.T. (2013). DExD/H-box RNA helicases in ribosome biogenesis. *RNA Biol.* 10, 4–18.
 - 133.Parsyan, A., Svitkin, Y., Shahbazian, D., Gkogkas, C., Lasko, P., Merrick, W.C., and Sonenberg, N. (2011). mRNA helicases: the tacticians of translational control. *Nat. Rev. Mol. Cell Biol.* 12, 235–245.
 - 134.Baldaccini, M., and Pfeffer, S. (2021). Untangling the roles of RNA helicases in antiviral innate immunity. *PLoS Pathog.* 17, e1010072.
 - 135.Sadler, A.J., and Williams, B.R.G. (2008). Interferon-inducible antiviral effectors. *Nat. Rev. Immunol.* 8, 559–568.
 - 136.Tijsterman, M., and Plasterk, R.H.A. (2004). Dicers at RISC; the mechanism of RNAi. *Cell* 117, 1–3.
 - 137.Bohnsack, K.E., Yi, S., Venus, S., Jankowsky, E., and Bohnsack, M.T. (2023). Cellular functions of eukaryotic RNA helicases and their links to human diseases. *Nat. Rev. Mol. Cell Biol.* 24, 749–769.
 - 138.Jackson, R.N., Lavin, M., Carter, J., and Wiedenheft, B. (2014). Fitting CRISPR-associated Cas3 into the helicase family tree. *Curr. Opin. Struct. Biol.* 24, 106–114.
 - 139.Domgaard, H., Cahoon, C., Armbrust, M.J., Redman, O., Jolley, A., Thomas, A., and Jackson, R.N. (2023). CasDinG is a 5'-3' dsDNA and RNA/DNA helicase with three accessory domains essential for type IV CRISPR immunity. *Nucleic Acids Res.*

<https://doi.org/10.1093/nar/gkad546>.

140. Schwarz, F.W., Tóth, J., van Aelst, K., Cui, G., Clausing, S., Szczelkun, M.D., and Seidel, R. (2013). The helicase-like domains of type III restriction enzymes trigger long-range diffusion along DNA. *Science* **340**, 353–356.
141. Bell, R.T., Juozapaitis, J., Songailiene, I., Sahakyan, H., Liegute, T., Makarova, K.S., Wolf, Y.I., Siksnys, V., and Koonin, E.V. (2023). Astonishing diversity and multifaceted biological connections of Type IV restriction-modification systems. *bioRxiv*, 2023.07.31.551357. <https://doi.org/10.1101/2023.07.31.551357>.
142. Rao, D.N., Dryden, D.T.F., and Bheemanaik, S. (2014). Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res.* **42**, 45–55.
143. Tesson, F., Planel, R., Egorov, A.A., Georjon, H., Vaysset, H., Brancotte, B., Neron, B., Mordret, E., Atkinson, G.C., Bernheim, A., et al. (2024). A Comprehensive Resource for Exploring Antiphage Defense: DefenseFinder Webservice, Wiki and Databases. *bioRxiv*, 2024.01.25.577194. <https://doi.org/10.1101/2024.01.25.577194>.
144. Millman, A., Melamed, S., Leavitt, A., Doron, S., Bernheim, A., Hör, J., Garb, J., Bechon, N., Brandis, A., Lopatina, A., et al. (2022). An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556–1569.e5.
145. Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., Doron, S., and Sorek, R. (2018). DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat Microbiol* **3**, 90–98.
146. Sather, L.M., Zamani, M., Muhammed, Z., Kearsley, J.V.S., Fisher, G.T., Jones, K.M., and Finan, T.M. (2023). A broadly distributed predicted helicase/nuclease confers phage resistance via abortive infection. *Cell Host Microbe* **31**, 343–355.e5.
147. Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., and Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183.
148. Thiaville, J.J., Kellner, S.M., Yuan, Y., Hutinet, G., Thiaville, P.C., Jumpathong, W., Mohapatra, S., Brochier-Armanet, C., Letarov, A.V., Hillebrand, R., et al. (2016). Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1452–E1459.
149. Payne, L.J., Hughes, T.C.D., Fineran, P.C., and Jackson, S.A. (2024). New antiviral defences are genetically embedded within prokaryotic immune systems. *bioRxiv*, 2024.01.29.577857. <https://doi.org/10.1101/2024.01.29.577857>.
150. O’Leary, N.A., Wright, M.W., Rodney Brister, J., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733.

151. Guy, C.P., and Bolt, E.L. (2005). Archaeal Hel308 helicase targets replication forks in vivo and in vitro and unwinds lagging strands. *Nucleic Acids Res.* 33, 3678–3690.
152. Johnson, S.J., and Jackson, R.N. (2013). Ski2-like RNA helicase structures: common themes and complex assemblies. *RNA Biol.* 10, 33–43.
153. Payne, L.J., Todeschini, T.C., Wu, Y., Perry, B.J., Ronson, C.W., Fineran, P.C., Nobrega, F.L., and Jackson, S.A. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res.* 49, 10868–10878.
154. Patel, I.R., Gangiredla, J., Mammel, M.K., Lampel, K.A., Elkins, C.A., and Lacher, D.W. (2018). Draft Genome Sequences of the *Escherichia coli* Reference (ECOR) Collection. *Microbiol Resour Announc* 7. <https://doi.org/10.1128/MRA.01133-18>.
155. Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., and Smith, D.B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* 46, D708–D717.
156. Calendar, R. *The Bacteriophages* (Springer US).
157. Davison, J. (2015). Pre-early functions of bacteriophage T5 and its relatives. *Bacteriophage* 5, e1086500.
158. Chaikerasitak, V., Nguyen, K., Khanna, K., Brilot, A.F., Erb, M.L., Coker, J.K.C., Vavilina, A., Newton, G.L., Buschauer, R., Pogliano, K., et al. (2017). Assembly of a nucleus-like structure during viral replication in bacteria. *Science* 355, 194–197.
159. Mendoza, S.D., Nieweglowska, E.S., Govindarajan, S., Leon, L.M., Berry, J.D., Tiwari, A., Chaikerasitak, V., Pogliano, J., Agard, D.A., and Bondy-Denomy, J. (2020). A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* 577, 244–248.
160. Fairman-Williams, M.E., Guenther, U.-P., and Jankowsky, E. (2010). SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* 20, 313–324.
161. Knizewski, L., Kinch, L.N., Grishin, N.V., Rychlewski, L., and Ginalski, K. (2007). Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches. *BMC Struct. Biol.* 7, 40.
162. Kennedy, M.A., Hosford, C.J., Azumaya, C.M., Luyten, Y.A., Chen, M., Morgan, R.D., and Stoddard, B.L. (2023). Structures, activity and mechanism of the Type IIS restriction endonuclease PqCI. *Nucleic Acids Res.* 51, 4467–4487.
163. Stokar-Avihail, A., Fedorenko, T., Hör, J., Garb, J., Leavitt, A., Millman, A., Shulman, G., Wojtania, N., Melamed, S., Amitai, G., et al. (2023). Discovery of phage determinants that confer sensitivity to bacterial immune systems. *Cell* 186, 1863–1876.e16.
164. Sasaki, T., Takita, S., Fujishiro, T., Shintani, Y., Nojiri, S., Yasui, R., Yonesaki, T., and

- Otsuka, Y. (2023). Phage single-stranded DNA-binding protein or host DNA damage triggers the activation of the AbpAB phage defense system. *mSphere*, e0037223.
165. Buss, J.A., Kimura, Y., and Bianco, P.R. (2008). RecG interacts directly with SSB: implications for stalled replication fork regression. *Nucleic Acids Res.* 36, 7029–7042.
 166. Bhattacharyya, B., George, N.P., Thurmes, T.M., Zhou, R., Jani, N., Wessel, S.R., Sandler, S.J., Ha, T., and Keck, J.L. (2014). Structural mechanisms of PriA-mediated DNA replication restart. *Proc. Natl. Acad. Sci. U. S. A.* 111, 1373–1378.
 167. Duckworth, A.T., Ducos, P.L., McMillan, S.D., Satyshur, K.A., Blumenthal, K.H., Deorio, H.R., Larson, J.A., Sandler, S.J., Grant, T., and Keck, J.L. (2023). Replication fork binding triggers structural changes in the PriA helicase that govern DNA replication restart in *E. coli*. *Nat. Commun.* 14, 2725.
 168. Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M., and Wood, T.K. (2010). Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* 1, 147.
 169. Umenhoffer, K., Fehér, T., Balikó, G., Ayaydin, F., Pósfai, J., Blattner, F.R., and Pósfai, G. (2010). Reduced evolvability of *Escherichia coli* MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. *Microb. Cell Fact.* 9, 38.
 170. Shen, L.L., and Pernet, A.G. (1985). Mechanism of inhibition of DNA gyrase by analogues of nalidixic acid: the target of the drugs is DNA. *Proc. Natl. Acad. Sci. U. S. A.* 82, 307–311.
 171. Booth, J.A., Špírek, M., Lobie, T.A., Skarstad, K., Krejci, L., and Bjørås, M. (2020). Antibiotic-induced DNA damage results in a controlled loss of pH homeostasis and genome instability. *Sci. Rep.* 10, 19422.
 172. Cabral, J.H.M., Jackson, A.P., Smith, C.V., Shikotra, N., Maxwell, A., and Liddington, R.C. (1997). Crystal structure of the breakage–reunion domain of DNA gyrase. *Nature* 388, 903–906.
 173. Kreuzer, K.N., and Cozzarelli, N.R. (1979). *Escherichia coli* mutants thermosensitive for deoxyribonucleic acid gyrase subunit A: effects on deoxyribonucleic acid replication, transcription, and bacteriophage growth. *J. Bacteriol.* 140, 424–435.
 174. Wohlkonig, A., Chan, P.F., Fosberry, A.P., Homes, P., Huang, J., Kranz, M., Leydon, V.R., Miles, T.J., Pearson, N.D., Perera, R.L., et al. (2010). Structural basis of quinolone inhibition of type IIA topoisomerases and target-mediated resistance. *Nat. Struct. Mol. Biol.* 17, 1152–1153.
 175. Laponogov, I., Sohi, M.K., Veselkov, D.A., Pan, X.-S., Sawhney, R., Thompson, A.W., McAuley, K.E., Fisher, L.M., and Sanderson, M.R. (2009). Structural insight into the quinolone–DNA cleavage complex of type IIA topoisomerases. *Nat. Struct. Mol. Biol.* 16, 667–669.
 176. Sugino, A., Higgins, N.P., Brown, P.O., Peebles, C.L., and Cozzarelli, N.R. (1978). Energy

- coupling in DNA gyrase and the mechanism of action of novobiocin. *Proc. Natl. Acad. Sci. U. S. A.* 75, 4838–4842.
177. Tsai, F.T., Singh, O.M., Skarzynski, T., Wonacott, A.J., Weston, S., Tucker, A., Paupit, R.A., Breeze, A.L., Poyser, J.P., O'Brien, R., et al. (1997). The high-resolution crystal structure of a 24-kDa gyrase B fragment from *E. coli* complexed with one of the most potent coumarin inhibitors, clorobiocin. *Proteins* 28, 41–52.
 178. Stubbe, J., and Kozarich, J.W. (1987). Mechanisms of bleomycin-induced DNA degradation. *Chem. Rev.* 87, 1107–1136.
 179. Verweij, J., and Pinedo, H.M. (1990). Mitomycin C: mechanism of action, usefulness and limitations. *Anticancer Drugs* 1, 5–13.
 180. Yoshizawa, S., Fourmy, D., and Puglisi, J.D. (1998). Structural origins of gentamicin antibiotic action. *EMBO J.* 17, 6437–6448.
 181. Mayo-Muñoz, D., Smith, L.M., Garcia-Doval, C., Malone, L.M., Harding, K.R., Jackson, S.A., Hampton, H.G., Fagerlund, R.D., Gummy, L.F., and Fineran, P.C. (2022). Type III CRISPR-Cas provides resistance against nucleus-forming jumbo phages via abortive infection. *Mol. Cell* 82, 4471–4486.e9.
 182. Kutter, E.M., and Wiberg, J.S. (1968). Degradation of cytosine-containing bacterial and bacteriophage DNA after infection of *Escherichia coli* B with bacteriophage T4D wild type and with mutants defective in genes 46, 47 and 56. *J. Mol. Biol.* 38, 395–411.
 183. Nechaev, S., and Severinov, K. (2003). Bacteriophage-induced modifications of host RNA polymerase. *Annu. Rev. Microbiol.* 57, 301–322.
 184. Echols, H., and Goodman, M.F. (1991). Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* 60, 477–511.
 185. Kreuzer, K.N., and Brister, J.R. (2010). Initiation of bacteriophage T4 DNA replication and replication fork dynamics: a review in the *Virology Journal* series on bacteriophage T4 and its relatives. *Viol. J.* 7, 358.
 186. Mosig, G. (1998). Recombination and recombination-dependent DNA replication in bacteriophage T4. *Annu. Rev. Genet.* 32, 379–413.
 187. Syeda, A.H., Hawkins, M., and McGlynn, P. (2014). Recombination and replication. *Cold Spring Harb. Perspect. Biol.* 6, a016550.
 188. Aguilera, A., and Gómez-González, B. (2008). Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.* 9, 204–217.
 189. Stekas, B., Yeo, S., Troitskaia, A., Honda, M., Sho, S., Spies, M., and Chemla, Y.R. (2021). Switch-like control of helicase processivity by single-stranded DNA binding protein. *Elife* 10. <https://doi.org/10.7554/eLife.60515>.

190. Bari, S.M.N., Chou-Zheng, L., Howell, O., Hossain, M., Hill, C.M., Boyle, T.A., Cater, K., Dandu, V.S., Thomas, A., Aslan, B., et al. (2022). A unique mode of nucleic acid immunity performed by a multifunctional bacterial enzyme. *Cell Host Microbe* 30, 570–582.e7.
191. Friedberg, E.C., Walker, G.C., Siede, W., and Wood, R.D. (2005). *DNA Repair and Mutagenesis* (American Society for Microbiology Press).
192. Keidel, A., Kögel, A., Reichelt, P., Kowalinski, E., Schäfer, I.B., and Conti, E. (2023). Concerted structural rearrangements enable RNA channeling into the cytoplasmic Ski238-Ski7-exosome assembly. *Mol. Cell* 83, 4093–4105.e7.
193. Büttner, K., Nehring, S., and Hopfner, K.-P. (2007). Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat. Struct. Mol. Biol.* 14, 647–652.
194. Devarkar, S.C., Schweibenz, B., Wang, C., Marcotrigiano, J., and Patel, S.S. (2018). RIG-I Uses an ATPase-Powered Translocation-Throttling Mechanism for Kinetic Proofreading of RNAs and Oligomerization. *Mol. Cell* 72, 355–368.e4.
195. Yasui, R., Washizaki, A., Furihata, Y., Yonesaki, T., and Otsuka, Y. (2014). AbpA and AbpB provide anti-phage activity in *Escherichia coli*. *Genes Genet. Syst.* 89, 51–60.
196. Lowey, B., Whiteley, A.T., Keszei, A.F.A., Morehouse, B.R., Mathews, I.T., Antine, S.P., Cabrera, V.J., Kashin, D., Niemann, P., Jain, M., et al. (2020). 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.
197. Richards, J.D., Johnson, K.A., Liu, H., McRobbie, A.-M., McMahon, S., Oke, M., Carter, L., Naismith, J.H., and White, M.F. (2008). Structure of the DNA repair helicase hel308 reveals DNA binding and autoinhibitory domains. *J. Biol. Chem.* 283, 5118–5126.
198. Barry Elizabeth R., and Bell Stephen D. (2006). DNA Replication in the Archaea. *Microbiol. Mol. Biol. Rev.* 70, 876–887.
199. Adelman, C.A., Lolo, R.L., Birkbak, N.J., Murina, O., Matsuzaki, K., Horejsi, Z., Parmar, K., Borel, V., Skehel, J.M., Stamp, G., et al. (2013). HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. *Nature* 502, 381–384.
200. Anand, R., Buechelmaier, E., Belan, O., Newton, M., Vancevska, A., Kaczmarczyk, A., Takaki, T., Rueda, D.S., Powell, S.N., and Boulton, S.J. (2022). HELQ is a dual-function DSB repair enzyme modulated by RPA and RAD51. *Nature* 601, 268–273.
201. Tafel, A.A., Wu, L., and McHugh, P.J. (2011). Human HEL308 localizes to damaged replication forks and unwinds lagging strand structures. *J. Biol. Chem.* 286, 15832–15840.
202. Yirmiya, E., Leavitt, A., Lu, A., Ragucci, A.E., Avraham, C., Osterman, I., Garb, J., Antine, S.P., Mooney, S.E., Hobbs, S.J., et al. (2024). Phages overcome bacterial immunity via diverse anti-defence proteins. *Nature* 625, 352–359.

203. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682.
204. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
205. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.
206. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539.
207. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
208. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35, 518–522.
209. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296.
210. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028.
211. Zulkower, V., and Rosser, S. (2020). DNA Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics* 36, 4350–4352.
212. Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345.
213. Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011). A modular cloning system for standardized assembly of multigene constructs. *PLoS One* 6, e16765.
214. Abedon, S.T. (2011). Lysis from without. *Bacteriophage* 1, 46–49.
215. Adler, B.A., Hessler, T., Cress, B.F., Lahiri, A., Mutalik, V.K., Barrangou, R., Banfield, J., and Doudna, J.A. (2022). Broad-spectrum CRISPR-Cas13a enables efficient phage genome editing. *Nat. Microbiol.* 7, 1967–1979.
216. Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* 152, 36–51.
217. Sanchez-Garcia, R., Gomez-Blanco, J., Cuervo, A., Carazo, J.M., Sorzano, C.O.S., and Vargas, J. (2021). DeepEMhancer: a deep learning solution for cryo-EM volume

- post-processing. *Commun Biol* 4, 874.
218. Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., and Ferrin, T.E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* 27, 14–25.
 219. Croll, T.I. (2018). ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* 74, 519–530.
 220. Afonine, P.V., Poon, B.K., Read, R.J., Sobolev, O.V., Terwilliger, T.C., Urzhumtsev, A., and Adams, P.D. (2018). Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr D Struct Biol* 74, 531–544.
 221. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682.
 222. Tuck, O.T., Hu, J.J., Lopez, S.C., Adler, B.A., O'Brien, C.E., Hsieh, K., Meredith, C., Loi, K.J., Yoon, P.H., Doherty, E.E., et al. (2025). Recurrent acquisition of nuclease-protease pairs in antiviral immunity. *Science*, eaea8769.
 223. Ryazansky, S., Kulbachinskiy, A., and Aravin, A.A. (2018). The expanded universe of prokaryotic Argonaute proteins. *MBio* 9. <https://doi.org/10.1128/mBio.01935-18>.
 224. Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., and Sorek, R. (2020). Bacterial Retrons Function In Anti-Phage Defense. *Cell* 183, 1551–1561.e12.
 225. Cui, Y., Dai, Z., Ouyang, Y., Fu, C., Wang, Y., Chen, X., Yang, K., Zheng, S., Wang, W., Tao, P., et al. (2025). Bacterial Hachiman complex executes DNA cleavage for antiphage defense. *Nat. Commun.* 16, 2604.
 226. Aravind, L., Iyer, L.M., and Burroughs, A.M. (2022). Discovering biological conflict systems through genome analysis: Evolutionary principles and biochemical novelty. *Annu. Rev. Biomed. Data Sci.* 5, 367–391.
 227. Makarova, K.S., Timinskas, A., Wolf, Y.I., Gussow, A.B., Siksnys, V., Venclovas, Č., and Koonin, E.V. (2020). Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antiviral defense. *Nucleic Acids Res.* 48, 8828–8847.
 228. Koonin, E.V., Makarova, K.S., Wolf, Y.I., and Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* 21, 119–131.
 229. Burroughs, A.M., Zhang, D., Schäffer, D.E., Iyer, L.M., and Aravind, L. (2015). Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res.* 43, 10633–10654.

230. Carfi, A., Pares, S., Duée, E., Galleni, M., Duez, C., Frère, J.M., and Dideberg, O. (1995). The 3-D structure of a zinc metallo-beta-lactamase from *Bacillus cereus* reveals a new type of protein fold. *EMBO J.* 14, 4914–4921.
231. Aravind, L. (1999). An evolutionary classification of the metallo- β -lactamase fold proteins. *In Silico Biol.* 1, 69–91.
232. Pettinati, I., Brem, J., Lee, S.Y., McHugh, P.J., and Schofield, C.J. (2016). The chemical biology of human metallo- β -lactamase fold proteins. *Trends Biochem. Sci.* 41, 338–355.
233. Pimentel, Z.T., and Zhang, Y. (2018). Evolution of the natural transformation protein, ComEC, in bacteria. *Front. Microbiol.* 9, 2980.
234. Silale, A., Lea, S.M., and Berks, B.C. (2021). The DNA transporter ComEC has metal-dependent nuclease activity that is important for natural transformation. *Mol. Microbiol.* 116, 416–426.
235. Mestre, M.R., Gao, L.A., Shah, S.A., López-Beltrán, A., González-Delgado, A., Martínez-Abarca, F., Iranzo, J., Redrejo-Rodríguez, M., Zhang, F., and Toro, N. (2022). UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids Res.* 50, 6084–6101.
236. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500.
237. Huang, P., Liu, J., Guo, L., Xu, D., Shen, L., Yan, P., Tong, C., Fei, W., Cheng, M., Li, Z., et al. (2025). The anti-phage mechanism of a widespread trypsin-MBL module. *bioRxiv*, 2025.09.03.673762. <https://doi.org/10.1101/2025.09.03.673762>.
238. Li, Y., Adams, D.W., Liu, H.W., Shaw, S.J., Uchikawa, E., Jaskólska, M., Stutzmann, S., Righi, L., Szczelkun, M.D., Blokesch, M., et al. (2025). Structure and activation mechanism of a lamassu phage defence system. *bioRxiv*, 2025.03.14.643221. <https://doi.org/10.1101/2025.03.14.643221>.
239. Haudiquet, M., Chakravarti, A., Zhang, Z., Ramirez, J.L., Herrero Del Valle, A., Olinares, P.D.B., Lavenir, R., Ahmed, M.A., de la Cruz, M.J., Chait, B.T., et al. (2025). Structural basis for Lamassu-based antiviral immunity and its evolution from DNA repair machinery. *bioRxiv*. <https://doi.org/10.1101/2025.04.02.646746>.
240. Riley, L.G., Ralston, G.B., and Weiss, A.S. (1996). Multimer formation as a consequence of separate homodimerization domains: the human c-Jun leucine zipper is a transplantable dimerization module. *Protein Eng.* 9, 223–230.
241. Burroughs, A.M., and Aravind, L. (2020). Identification of uncharacterized components of prokaryotic immune systems and their diverse eukaryotic reformulations. *J. Bacteriol.* 202. <https://doi.org/10.1128/JB.00365-20>.
242. Yu, J.W., Jeffrey, P.D., Ha, J.Y., Yang, X., and Shi, Y. (2011). Crystal structure of the

- mucosa-associated lymphoid tissue lymphoma translocation 1 (MALT1) paracaspase region. *Proc. Natl. Acad. Sci. U. S. A.* *108*, 21004–21009.
243. Zhou, H., Du, M.-Q., and Dixit, V.M. (2005). Constitutive NF-kappaB activation by the t(11;18)(q21;q21) product in MALT lymphoma is linked to deregulated ubiquitin ligase activity. *Cancer Cell* *7*, 425–431.
 244. Jongeneel, C.V., Bedinger, P., and Alberts, B.M. (1984). Effects of the bacteriophage T4 dda protein on DNA synthesis catalyzed by purified T4 replication proteins. *J. Biol. Chem.* *259*, 12933–12938.
 245. Van Opdenbosch, N., and Lamkanfi, M. (2019). Caspases in cell death, inflammation, and disease. *Immunity* *50*, 1352–1364.
 246. Heutinck, K.M., ten Berge, I.J.M., Hack, C.E., Hamann, J., and Rowshani, A.T. (2010). Serine proteases of the human immune system in health and disease. *Mol. Immunol.* *47*, 1943–1955.
 247. Hu, C., van Beljouw, S.P.B., Nam, K.H., Schuler, G., Ding, F., Cui, Y., Rodríguez-Molina, A., Haagsma, A.C., Valk, M., Pabst, M., et al. (2022). Craspase is a CRISPR RNA-guided, RNA-activated protease. *Science*, eadd5064.
 248. Strecker, J., Demircioglu, F.E., Li, D., Faure, G., Wilkinson, M.E., Gootenberg, J.S., Abudayyeh, O.O., Nishimasu, H., Macrae, R.K., and Zhang, F. (2022). RNA-activated protein cleavage with a CRISPR-associated endopeptidase. *Science* *378*, 874–881.
 249. Steens, J.A., Bravo, J.P.K., Salazar, C.R.P., Yildiz, C., Amieiro, A.M., Köstlbacher, S., Prinsen, S.H.P., Andres, A.S., Patinios, C., Bardis, A., et al. (2024). Type III-B CRISPR-Cas cascade of proteolytic cleavages. *Science* *383*, 512–519.
 250. Ershova, A.S., Karyagina, A.S., Vasiliev, M.O., Lyashchuk, A.M., Lunin, V.G., Spirin, S.A., and Alexeevski, A.V. (2012). Solitary restriction endonucleases in prokaryotic genomes. *Nucleic Acids Res.* *40*, 10107–10115.
 251. Fiddaman, S.R., Vinkler, M., Spiro, S.G., Levy, H., Emerling, C.A., Boyd, A.C., Dimopoulos, E.A., Vianna, J.A., Cole, T.L., Pan, H., et al. (2022). Adaptation and cryptic pseudogenization in penguin Toll-like receptors. *Mol. Biol. Evol.* *39*, msab354.
 252. Béchon, N., Tal, N., Stokar-Avihail, A., Savidor, A., Kupervaser, M., Melamed, S., Amitai, G., and Sorek, R. (2024). Diversification of molecular pattern recognition in bacterial NLR-like proteins. *Nat. Commun.* *15*, 9860.
 253. Essuman, K., Milbrandt, J., Dangl, J.L., and Nishimura, M.T. (2022). Shared TIR enzymatic functions regulate cell death and immunity across the tree of life. *Science* *377*, eabo0001.
 254. Hogrel, G., Guild, A., Graham, S., Rickman, H., Grüşchow, S., Bertrand, Q., Spagnolo, L., and White, M.F. (2022). Cyclic nucleotide-induced helical structure activates a TIR immune effector. *Nature*. <https://doi.org/10.1038/s41586-022-05070-9>.

255. Morehouse, B.R., Yip, M.C.J., Keszei, A.F.A., McNamara-Bordewick, N.K., Shao, S., and Kranzusch, P.J. (2022). Cryo-EM structure of an active bacterial TIR-STING filament complex. *Nature* 608, 803–807.
256. Asplund-Samuelsson, J., Bergman, B., and Larsson, J. (2012). Prokaryotic caspase homologs: phylogenetic patterns and functional characteristics reveal considerable diversity. *PLoS One* 7, e49888.
257. Uren, A.G., O'Rourke, K., Aravind, L.A., Pisabarro, M.T., Seshagiri, S., Koonin, E.V., and Dixit, V.M. (2000). Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* 6, 961–967.
258. Koonin, E., and Aravind, L. (2002). Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ.* 9, 394–404.
259. Wilson, K.P., Black, J.A., Thomson, J.A., Kim, E.E., Griffith, J.P., Navia, M.A., Murcko, M.A., Chambers, S.P., Aldape, R.A., and Raybuck, S.A. (1994). Structure and mechanism of interleukin-1 beta converting enzyme. *Nature* 370, 270–275.
260. McIlwain, D.R., Berger, T., and Mak, T.W. (2013). Caspase functions in cell death and disease. *Cold Spring Harb. Perspect. Biol.* 5, a008656.
261. Johnson, A.G., Wein, T., Mayer, M.L., Duncan-Lowey, B., Yirmiya, E., Oppenheimer-Shaanan, Y., Amitai, G., Sorek, R., and Kranzusch, P.J. (2022). Bacterial gasdermins reveal an ancient mechanism of cell death. *Science* 375, 221–225.
262. Wein, T., Millman, A., Lange, K., Yirmiya, E., Hadary, R., Garb, J., Melamed, S., Amitai, G., Dym, O., Steinruecke, F., et al. (2025). CARD domains mediate anti-phage defence in bacterial gasdermin systems. *Nature*, 1–8.
263. Rousset, F., Osterman, I., Scherf, T., Falkovich, A.H., Leavitt, A., Amitai, G., Shir, S., Malitsky, S., Itkin, M., Savidor, A., et al. (2025). TIR signaling activates caspase-like immunity in bacteria. *Science* 387, 510–516.
264. Enari, M., Sakahira, H., Yokoyama, H., Okawa, K., Iwamatsu, A., and Nagata, S. (1998). A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature* 391, 43–50.
265. Eddy, S.R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7, e1002195.
266. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534.
267. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.

268. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 52, W78–W82.
269. Acids research, N., and 2018 Database resources of the national center for biotechnology information. academic.oup.com.
270. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* 32, e4792.
271. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175.
272. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
273. Revell, L.J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12, e16505.
274. Morel, B., Kozlov, A.M., Stamatakis, A., and Szöllősi, G.J. (2020). GeneRax: A tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* 37, 2763–2774.
275. Edgar, R.C. (2021). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* 13. <https://doi.org/10.1038/s41467-022-34630-w>.
276. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
277. Holm, L. (2022). Dali server: structural unification of protein families. *Nucleic Acids Research* 50, W210–W215.
278. Maffei, E., Shaidullina, A., Burkolter, M., Heyer, Y., Estermann, F., Druelle, V., Sauer, P., Willi, L., Michaelis, S., Hilbi, H., et al. (2021). Systematic exploration of *Escherichia coli* phage–host interactions with the BASEL phage collection. *PLoS Biol.* 19, e3001424.
279. Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* 403, 335–338.
280. Weaver, S.D., DeRosa, C.M., Schultz, S.R., and Champion, M.M. (2023). PrIntMap-R: An online application for intraprotein intensity and peptide visualization from bottom-up proteomics. *J. Proteome Res.* 22, 432–441.